Ad hoc and general-purpose corpus construction from web sources

Adrien Barbaresi

THÈSE

en vue de l'obtention du grade de

Docteur de l'Université de Lyon, délivré par l'École Normale Supérieure de Lyon

Discipline : Linguistique

Laboratoire ICAR

École Doctorale LETTRES, LANGUES, LINGUISTIQUE, ARTS

présentée et soutenue publiquement le 19 juin 2015

par Monsieur Adrien Barbaresi

Construction de corpus généraux et spécialisés à partir du web

---

Directeur de thèse : M. Benoît HABERT

Devant la commission d'examen formée de :

M. Benoît HABERT, ENS Lyon, Directeur

M. Thomas LEBARBÉ, Université Grenoble 3, Examinateur

M. Henning LOBIN, Université de Gießen, Rapporteur

M. Jean-Philippe MAGUÉ, ENS Lyon, Co-Encadrant

M. Ludovic TANGUY, Université Toulouse 2, Rapporteur

Ad hoc and general-purpose corpus construction from web sources

Adrien Barbaresi

2015

École Normale Supérieure de Lyon (Université de Lyon)
ED 484 3LA
ICAR lab

Thesis Committee:

Benoît	**Habert**	**ENS Lyon**	**advisor**
Thomas	Lebarbé	University of Grenoble 3	*chair*
Henning	Lobin	University of Gießen	*reviewer*
Jean-Philippe	Magué	ENS Lyon	*co-advisor*
Ludovic	Tanguy	University of Toulouse 2	*reviewer*

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics

Acknowledgements

Many thanks to

Benoît Habert and Jean-Philippe Magué
Henning Lobin
Serge Heiden, Matthieu Decorde, Alexei Lavrentev, Céline Guillot, and the people at ICAR
Alexander Geyken, Lothar Lemnitzer, Kay-Michael Würzner, Bryan Jurish, and the team at the *Zentrum Sprache* of the BBAW
my fellow former members of the board of ENthèSe
Felix Bildhauer and Roland Schäfer

the liberalities of my PhD grant and the working conditions I benefited from[1]

the S.
my friends and family

Eva Brehm-Jurish for helping my English sound better
Brendan O'Connor for his inspirational LaTeX preamble

[1] PhD grant at the ENS Lyon, internal research grant at FU Berlin, DFG-funded research grants at the BBAW. I am grateful to the FU Berlin and the BBAW for the substantial computational resources they provided.

Contents

List of Tables

List of Figures

Abstract

This thesis introduces theoretical and practical reflections on corpus linguistics, computational linguistics, and web corpus construction. More specifically, two different types of corpora from the web, specialized (*ad hoc*) and general-purpose, are presented and analyzed, including suitable conditions for their creation.

At the beginning of the first chapter the interdisciplinary setting between linguistics, corpus linguistics, and computational linguistics is introduced. The frame of the thesis is established in an interdisciplinary context, between linguistics and computational linguistics, with excursions into applied approaches.

Then, the notion of corpus is put into focus. In a synchronic perspective, the current need for linguistic evidence encompassing several disciplines such as theoretical linguistics or information retrieval is illustrated with several usage scenarios. Existing corpus and text definitions are discussed, the traditional notion of a corpus as a well-organized collection of carefully selected texts or text samples is introduced. Corpus and text typologies cement a real difference between general and specialized corpora, as well as different organizational constraints such as production date as well as the type of texts.

In a historical perspective, several milestones of corpus design are presented, from pre-digital corpora at the end of the 1950s to web corpora in the 2000s and 2010s. Three main phases are distinguished in this evolution, first the age of copy typing and establishment of the scientific methodology and tradition regarding corpora, second the age of digitized text and further development of corpus linguistics, and third the arrival of web data and “opportunistic" approaches among researchers. The empirical trend of corpus linguistics emerging in the 1980s is analyzed as being part of a global trend fostered by technological evolutions toward instrumentalized linguistics and big data. The statistical view on language and the belief among corpus linguists that language patterns are more interesting than the discovery or description of rules have an influence on corpus construction, all the more since text material becomes (seemingly) easily accessible on the Web.

The continuities and changes between the linguistic tradition, i.e. existing “reference" digital corpora, and web native corpora are exposed. These concern methodological issues in corpus linguistics such as text typology and representativeness, as well as practical challenges related for instance to metadata and suitable toolchains. Reasons to advocate for the “Web for corpus" approach rather than the “Web as corpus" standpoint are given, and new challenges for corpus building such as less-resourced languages or language documentation are presented.

The first chapter is concluded with considerations on the rise of web corpus linguistics, and

the consequent changes in corpus construction. As unquestioned aspects of corpus linguistics practice are summarized, challenges addressed in this thesis are highlighted: data quality and exploitability; accessible, independent, and practicable web corpus gathering, as well as approaches crossing several disciplines and research traditions in France, Germany, and elsewhere.

In the second chapter, methodological insights on automated text scrutiny in computer science, computational linguistics and natural language processing are presented.

The state of the art on text quality assessment and web text filtering is described in order to exemplify current interdisciplinary research trends on web texts. Frequently occurring text quality issues are presented, such as machine-generated text, multiple languages, or spam. Current solutions to these problems are explored, as proxies have to be found for phenomena which strike the human eye but which machines fail to detect.

Readability studies and automated text classification are used as a paragon of methods to find salient features in order to grasp text characteristics. As research on readability is also interdisciplinary, the field and its salient challenges are summarized. Several methods are compared, from theoretical to industrial, including strictly applied approaches. Common denominators are listed, such as preparation of the classification and training data. Efficient criteria are summarized, such as the surprising efficiency of surface features, and feature selection techniques.

Last, text visualization exemplifies corpus processing in the digital humanities framework. The difference between global and local visualization techniques is discussed. Several historic examples of corpus visualization are given. Then, the interest of visual analytics and information visualization, a disciplinary field of computer science, is presented from the angle of linguistic applications.

As a conclusion, guiding principles for research practice are listed, and reasons are given to find a balance between quantitative analysis and corpus linguistics, in an environment which is spanned by technological innovation and artificial intelligence techniques.

Third, current research on web corpora is summarized. The chapter opens with notions of “web science" such as a definition of web crawling and an overview of the constraints of “offline" web corpora.

Then, I examine the issue of data collection, more specifically in the perspective of URL seeds, both for general and for specialized corpora. Problems for existing approaches concern stability and exhaustivity of procedures. I distinguish two main approaches to web document retrieval: restricted retrieval, where documents to be retrieved are listed or even known in advance, and web crawling. I show that the latter case should not be deemed too complex for linguists, by summarizing different strategies to find new documents, and discussing their advantages and limitations. Finally, ways to target small fractions of the Web and afferent issues are described.

In a further section, the notion of web corpus preprocessing is introduced and salient steps are discussed, such as filtering, cleaning, inclusion into the corpus and controls of the corpus. The impact of the preprocessing phase on research results is assessed, both from a quantitative and qualitative point of view, with practical linguistic examples. To conclude, I explain why the importance of preprocessing should not be underestimated and why it is an important task for linguists to learn new skills in order to confront the whole data gathering and preprocessing phase.

As web crawling is only seemingly simple, issues are raised for the actual corpus work described in this thesis, most importantly for assessing the relevance of web data for corpus research, gathering large quantities of text or particular fractions of the Web on a budget, and making the whole procedure reproducible.

In consequence, I present my work on web corpus construction in the fourth chapter, with two types of end products, specialized and even niche corpora on the one hand, and general-purpose corpora on the other hand.

My analyses concern two main aspects, first the question of corpus sources (or *prequalification*), and secondly the problem of including valid, desirable documents in a corpus (or document *qualification*).

First, I show that it is possible and even desirable to use sources other than just search engines as state of the art, and I introduce a light scout approach along with experiments to prove that a preliminary analysis and selection of crawl sources is possible as well as profitable.

Second, I perform work on document selection, in order to enhance web corpus quality in general-purpose approaches, and in order to perform a suitable quality assessment in the case of specialized corpora. I show that it is possible to use salient features inspired from readability studies along with machine learning approaches in order to improve corpus construction processes. To this end, I select a number of features extracted from the texts and tested on an annotated sample of web texts.

Last, I present work on corpus visualization consisting of extracting certain corpus characteristics in order to give indications on corpus contents and quality.

Chapter *1*

Corpus design before and after the emergence of web data

1.1 Prologue: “A model of Universal Nature made private" – Garden design and text collections

Si hortum in bibliotheca habes, nihil deerit.[1]
Marcus Tullius Cicero, *Ad familiares* IX, 4, to Varro.

The quoted aphorism by Cicero exemplifies that gardens are usually closely tied to human culture and knowledge, for example through the question of nature and culture. Gardens are strictly speaking cultivated places, which offer a sample of nature as seen, conceived, symbolized, exemplified, studied, or refined by man. There are many *topoi* linked to the idea of a garden. In German humanities, the figure of the gardener seen as a scientist and vice-versa has been in existence at least since Herder (Burbulla, 2011), and there are numerous examples of the intricate ties between science and garden design, be it through the lives of scientists, for example Leibniz and the Herrenhausen gardens, or as a frequent motive in literature, noticeable for instance in Goethe's *Elective Affinities*.

The topic seems to be rich enough that according to Fischer, Remmert, and Wolschke-Bulmahn (2011), an extensive, systematic study of the historical context linking knowledge, science (especially experimental sciences), and gardens remains to be performed.

Tabarasi-Hoffmann (2011) explains that Francis Bacon's expression of “a model of Universal Nature made private" is typical for the interest of the proponents of scientific method and empiricism for gardens and zoological gardens, which are seen as laboratories, experimental fields, paragons of the control of science and technological subjection of nature. On the sideline, nature enhanced through science also seems to be considered in an eschatological perspective, with the myth of the tree of knowledge in the background.

To come back to Cicero, in the case of a library of language data, gardens are more than just a resting place for the eye or solace for the mind, whether considered to be a world *in nuce*, i.e. in a nutshell, or a so-called “cabinet of rarities". As an introduction to the matter of corpus construction I would like to suggest that gardens and text collections are comparable in their purpose of being an extract, or even sample of nature which has been manipulated by man, and fosters and influences the course of his reflexions.

Concrete points of comparison Thus, collections of language samples can be compared to garden design in a way. Although the latter may be considered to be an art, the same principles apply in the sense that both are necessarily limited collections meant to thrive in their artificiality or, much more often, to be a sample of nature and to offer a certain variety in a smaller and more contained space than would otherwise be the case in the open.

The most striking common aspect of landscape architecture and text collections is probably the fact that they are formed by strong design decisions, which can be inspired by long-lived traditions, as well as by resolute concepts of order, system, even orderliness, or, conversely, contingency and serendipity.

Additionally, there are a number of other characteristics making them comparable. Both can be secluded places hidden behind walls, or open to cross-fertilization. Both may include or exclude the outside world, for instance include it as a realistic, idealized or partial reduction,

[1] Literal translation: If you have a garden in your library, nothing will be lacking.
Liberal translation: If you have a garden and a library, you have everything you need.

or exclude it because of its radically different nature, by gathering samples which have been collected in remote locations.

A few types of garden Garden design is probably much older than reasoning about language, and its history still fosters prolific research. The following garden types, unfairly simplified without paying respect to their evolution in the course of history, are listed because of their distinctive characteristics.[2]

The French formal garden imposes order on nature. It is laid out upon a geometric plan with a strong emphasis on axis and perspective. The trees are carefully trimmed, with tops at a set height.

The English garden is supposed to be an artificial image of nature which is meant to be as close to the original as possible.

Japanese gardens are miniature idealized landscapes, with no fixed allowances of space. Smaller gardens are often designed to incorporate the view of features outside the garden, such as hills, trees or temples, as part of the view. This makes the garden seem larger than it really is.

Thus, design decisions play a paramount role in the arrangement of gardens. The notion of point of view is also highly relevant, with particular emphasis on general layout and specific experiences of visitors strolling along the paths. In French gardens, an overlook, for instance in form of a terrace, is possible and even desired. Chinese gardens are designed to be seen from the inside, from the buildings, galleries and pavilions in the center of the garden. Later Japanese gardens are designed to be seen from the outside, or from a path winding through the garden.

Last, the notion of variety is also present to a variable extent among garden cultures, it is sometimes considered to be highly relevant, and sometimes secondary with respect to the success of a garden.

The text in linguistics: herbarium or garden? Comparisons of examples and text scrutiny in linguistics with nature exist. However, there seems to be a strong tendency towards considering the text collections used in linguistics as a herbarium rather than a garden:

> "Faire de la linguistique sur des textes, c'est faire de la botanique sur un herbier, de la zoologie sur des dépouilles d'animaux plus ou moins conservées."[3]
> Discours de réception de Charles de Tourtoulon[4], *Académie des sciences, agriculture, arts et belles lettres*, Aix-en-Provence, 1897.

As a matter of fact, Tourtoulon was far more interested in spoken language than in written texts, all the more since text publishing in rural southern France at the end of the 19th century probably did not give a realistic image of the way people spoke in everyday life.

The image of the herbarium is still used one century later in French linguistics:

[2] https://en.wikipedia.org/wiki/Category:Types_of_garden_by_country_of_origin

[3] "To do linguistics on texts is to do botanics on a herbarium, and zoology on remains of more or less well-preserved animals."

[4] Charles-Jean-Marie de Tourtoulon, Carles de Tortolon in langue d'oc, (1836-1913) has, among other things, contributed to establish a map of the distribution of the langue d'oc based on empirical investigations.

> “À dire vrai, l’exemple n’est pas dans une grammaire comme une plante dans un champ, mais comme une plante dans un herbier."[5] (Auroux, 1998, p. 185)

Auroux continues by saying that the plant in a linguists’ herbarium is an occurrence representing its type, a proto-type of language use. In this particular case, Auroux speaks of example sentences and not full texts, in a context which is closer to the *topos* of the medieval *hortus conclusus*, a secluded garden which presents an ordered version of the outside world, and naturally tends to be used for classification and inventory purposes. Nonetheless, this case highlights the conceptual proximity between linguistic and botanic study.

As shown in this prologue, the images of nature given by a few types of gardens as well as the reduction to a herbarium can yield useful comparisons with corpus design and perspectives on linguistics. According to Chevalier (1997), the image of an herbarium is a paragon of the French theoretical, abstract perspective on linguistics.

Since my thesis touches on this comparison in two places, I will return to this image in the course of the work. First, to understand how gardens were made available instead of mere herbariums and what it changed for linguists. Second, to explain and tackle some of the challenges raised by a greater language sample availability than ever before, most notably how a limited garden can display a reasonable image of nature in its diversity and constant evolution, with its vividness and potential for surprise.

1.2 Introductory definitions and typology: corpus, corpus design and text

1.2.1 What “computational" is to linguistics

The terms “Computational Linguistics" and “Natural Language Processing" are found in the literature embracing a vast array of research topics, techniques, usage scenarios, and theoretical backgrounds having to do with the use of computers to (among other things) analyze, study, replicate and/or explain language and language phenomena. The use of NLP currently seems to be more frequent in the United States, whereas in France or Germany for instance the notions of “linguistique informatique/computationelle" and “Computerlinguistik" clearly refer to Computational Linguistics. In France, the denomination “traitement automatique du langage naturel" distinguishes the applied dimension from general computational linguistics.

Most of the time, NLP and computational linguistics are practically used as synonyms, although they may refer to different scientific traditions and diverging research goals, as Kay (2005) explains:

> “Computational linguistics is not natural language processing. Computational linguistics is trying to do what linguists do in a computational manner, not trying to process texts, by whatever methods, for practical purposes. Natural language processing, on the other hand, is motivated by engineering concerns." (Kay, 2005, p. 429)

The following work falls resolutely on the linguistic side. It does not primarily deal with engineering concerns, although it may at some point become tangential to them. It rather

[5] “In fact, the example is not in a grammar like a plant in a field, but like a plant in a herbarium."

tackles research issues related to corpus construction on the web, which may seem technical or motivated by practical issues. However, one of the points which I want to highlight in this thesis is that these issues are actually crucial since they impact further linguistic research on corpora.

The disciplinary fields of computational and corpus linguistics have not been clearly delimited and categorized yet, and a proper description of them falls beyond the scope of this work. I will start by describing what a corpus is or is believed to be and how the evolution of scientific interests affects it.

1.2.2 What theory and field are to linguistics

Linguistics and computational linguistics are disciplines centered rather on the side of the so-called “armchair linguist" (Fillmore, 1992) than on an experimental side since the 1950s. The fact that quite theoretical descriptions of phenomena prevailed has had an influence over the evolution of these disciplines.

At first, the disciplinary field of academic linguistics seems to have been mostly defined by the theorization of phenomena rather than the collection of empirical knowledge. The 1960s saw the beginnings of the institutionalization of field linguistics on a separate basis, which developed independently and in part against the primacy of introspection and theory, which is targeted by the ironical denomination of “armchair" linguistics. Since then, both traditions have mostly evolved in parallel.[6]

According to Philippe Blanchet, who puts a strong emphasis on field linguists and thus cannot be expected to be neutral, the divergence concerns the epistemological models. One tradition integrates criteria from the sciences, speaks of a “scientific theory" of language, uses a “hypothetico-deductive method applied to the atomistic analysis of an ‘internal mechanic’"[7] and aims to look for logical-mathematical objectivity and predictability. The other tradition includes criteria related to specific, described human practices, uses an “empirical-inductive method of observation in context"[8] and presents a nuanced interpretation (Blanchet, 2003, p. 288).

Fillmore (1992), by comparing the stereotypes of the “armchair" and the field linguists, develops a similar view on linguistics. One may criticize the lack of neutrality of these assumptions, but the interesting fact is precisely that there are researchers on the side of an empiricist tradition who feel that the prevalence of theory has to be compensated by setting things straight, or even a bit straighter than straight.

Additionally, the primacy of theory is also a French phenomenon encouraged by social structures and a global scientific tradition (Bergounioux, 1992).[9] All in all, the primacy of

[6] “L’objet de la linguistique, du structuralisme saussurien au générativisme, a donc été construit en accordant la priorité à l’invariant, et à la dissociation entre phénomènes ‘linguistiques’ d’une part (le code) et phénomènes communicationnels et socioculturels d’autre part (les usages des codes)." (Blanchet, 2003, p. 287)

[7] “Méthode hypothético-déductive appliquée à l’analyse atomisante d’une ‘mécanique interne’"

[8] “Méthode empirico-inductive d’observation en contexte"

[9] “Le marché de la linguistique en France, centralisé par le quasi-monopole de Paris sur la circulation des biens symboliques et dominé par les études littéraires dans l’organisation universitaire, favorise objectivement la recherche abstraite, universalisante ou formalisante, au détriment des pratiques de recension descriptives, autrement dit, ceux qui travaillent en bibliothèque plutôt que sur le terrain. La présence dans les bureaucraties d’administration de la science, par exemple, est plus favorable à une carrière dans l’enseignement supérieur que l’enquête." (Bergounioux, 1992, p. 18)

introspection and the notion of an expert, trained linguist reasoning on language may have more momentum and inertia in France than in Germany for instance.

From the perspective of the philosopher of technology Gilbert Hottois, one could add a subtrend on the side of theory, the notion of operativity. Hottois sees in generative grammar the "most complete expression of the replacement of linguistic and theoretical essence of mankind by its other, of operative nature."[10]. The distinctions above may seem too clear-cut, and the description by Hottois is somewhat far-fetched, since generative grammar can be seen as an heir of 1930s linguistics, which put an emphasis on field observation. Nonetheless, Hottois' statement about the growing importance of being operative as an increasingly popular scientific paradigm may be considered as valid in certain cases. Similarly, the distinctions by Fillmore and Blanchet exemplify a wish to anchor two different perspectives on linguistics.

There are indeed cases where linguistics tends to become more of an applied discipline, where explanatory power and direct applications are powerful arguments for both research theory and practice. Additionally, there are research trends going beyond the field of linguistics which also bring a change of perspective towards more applicative science (see p. 31).

1.2.3 Why use a corpus? How is it defined? How is it built?

1.2.3.1 Potential usage scenarios

According to (Baroni & Ueyama, 2006), in a seminal article about general web corpora, corpora are about bringing "actual usage evidence". The authors distinguish four main cases where such evidence is needed (Baroni & Ueyama, 2006, p. 31):

- theoretical and applied linguistic questions
- simulations of language acquisition
- lexicography
- a large number of tasks in natural language processing

In fact, several usage scenarios and related research communities can be distinguished. The few mains types described below are not supposed to be perfect matches for real life researchers, they are rather to be seen as an illustration of the various research interests and practices behind corpora. In the following I will distinguish four different cases in linguistics, computational linguistics, and beyond: theoretical linguists, applied/corpus/experimental linguists, computational linguists and NLP, information retrieval and information extraction, and finally I give an outlook into other disciplines.

Theoretical linguists First of all, there are theoretical linguists who work with corpora because they want to find actual facts to support the hypotheses they construct about the functioning of language or particular language phenomena. They may also test their opinions on actual corpora, in the sense that the corpus "enhances our understanding of the workings of language, and forces us to rethink our own subjective perceptions." (Rundell & Stock, 1992, p. 29)

[10] "La grammaire générative est l'expression la plus achevée du remplacement de l'essence langagière et théorique de l'homme par son autre opératoire." (Hottois, 1984, p. 61)

One may take word order in German as an example, as a typical case of language evolution, where traditional grammarians stipulate that a conjugated verb is to be found at the end of subordinates. There are a number of cases where this is false, such as causal *weil*-subordinates where the verb comes in second position as if it were the main phrase of the sentence.

Theoretical linguists first have to become aware of such phenomena, which mostly happens by a sort of day-to-day serendipity and/or by the random exploration of corpora, and then they have to formulate a hypothesis and to find contradictory cases to validate it, so that a typical extract of a theoretical linguist's article may look like the following:

(1) Könnten Sie sich bitte woanders hinsetzen? *Weil* das *sind* unsere Plätze.[11]

Finding examples for verb-second subordinate structures ("weil + V2") is relatively easy in spoken language, but much more difficult in written text. Since the probability of finding it by chance, just by listening to people talk or reading texts, is relatively low, one may prefer to query a text corpus to look for this particular structure and then deduce that, for example, there are "strong", classical usages of *weil* and others where the status of the causal particle is weakened so that it turns from a subjunction into a conjunction comparable to *denn* and/or introduces an adverbial phrase (Freywald, 2009).

A large fringe of research in lexicography is roughly comparable to theoretical linguistics since it uses corpora to look for attested language usages, for instance to complement a given meaning in a definition. The main difference is probably that most lexicographers aim for the largest possible coverage, while theoretical linguists are more likely to put an emphasis on data selection.

To conclude, theoretical linguists use corpora either for hypothesis testing or in an exploratory fashion: two different ways to see corpora, although there might be an hermeneutic circle between theory and practice. For a discussion on research methodology, see p. 24.

Applied/corpus/experimental linguists Applied linguists deal with resources in an extensive way. It is not only about handpicking examples, it is about proving as well as finding theories by using quantitative methods on corpora.

A current trend is to look at the way brains learn language in a statistical way, looking for patterns (see p. 41). One example would be the study of word collocations.

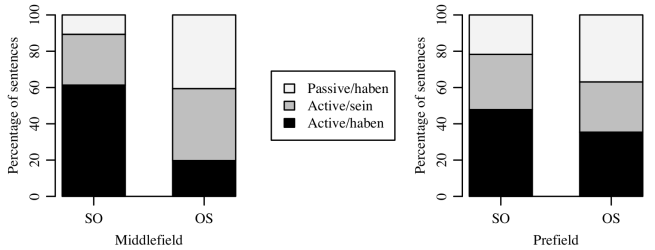


Figure 5: Distribution of verb properties in sentences with a dative object according to order (SO vs. OS) and position (middlefield vs. prefield). ?

Figure 1.1: Results of a corpus study presented in a graphical fashion

[11]The author who mentions this example (Freywald, 2009) does not cite any source, it is thus unclear whether the sentence has been specially forged or not.

Figure 1.1 shows results obtained by Bader and Häussler (2010) during a corpus study on word order in German. The bars represent the distribution of the objects of study in a corpus. An introduction of such results in form of a table is also standard, and tables are also used in the same article.

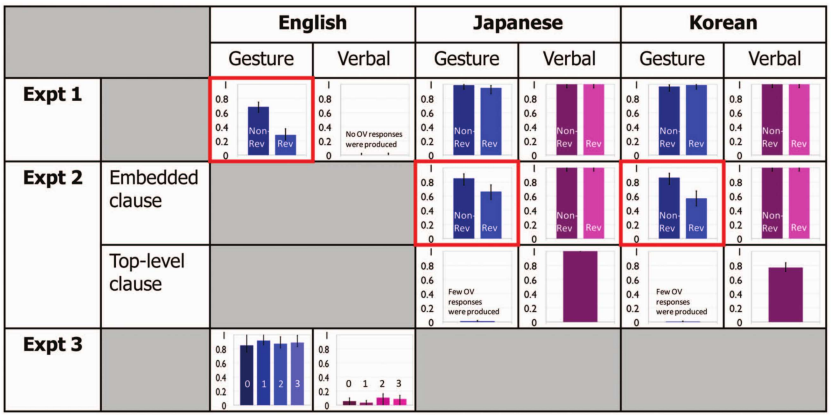


Fig. 3. Summary of results for English, Japanese, and Korean speakers in Experiments 1 through 3. For all experiments, the graphs show the proportion of patient-before-action (object-before-verb, or OV) productions; results for gestured responses are in blue, on the left, and results for verbal responses are in pink, on the right. For Experiments 1 and 2, responses are shown separately for nonreversible ("Non-Rev") and reversible ("Rev") events; for Experiment 2, the top row shows results for embedded events, and the bottom row shows results for top-level events. The graphs for Experiment 3 show the proportion of OV responses as a function of the number of features on the patient (0, 1, 2, or 3). The gesture patterns for reversible and nonreversible events that provide critical evidence in support of the noisy-channel hypothesis are highlighted by the red outlines. Error bars represent 95% confidence intervals.

Figure 1.2: Results of a comparative psycholinguistic experiment

Figure 1.2 shows the results of a psycholinguistic study performed by Gibson et al. (2013) on word order. The answers of the response panel are then used to confirm, invalidate or derive theories on language. In the particular case of experimental psycholinguistics, the corpus is a means and not an end in itself, since it is used to extract sentences which are eventually modified before being used as experimental material.

In both cases, text quantity is a relevant factor in order to generate statistically reliable results, whereas for theoretical linguists quantity can only be secondary, either in order to have a bigger reservoir of examples at one's command or to increase the chances of finding evidence, for instance.

Computational linguistics and NLP The quantity of available text is also relevant to most of today's researchers and engineers in natural language processing, who are interested in developing language-based applications. Most of them use statistical and/or machine-learning methods rather than rule-based ones. In that sense, this particular field has been a precursor of a revalorization of empiricism in (computational) linguistics (see p. 21).

Figure 1.3 (Crowgey & Bender, 2011) exemplifies how sentences from a corpus may be used to test assumptions regarding the construction and application of grammars, i.e. formalized models of language. The hierarchical structure is the parse (syntactic structure) of a sentence, whereas the cells contain lexico-syntactic information which are derived from other processing steps and which may be used to extract the sentence structure.

Figure 1.4 illustrates another tendency in computational linguistics and language modeling. In fact, the study by Bader and Häussler (2010), which I also mention above in the context of corpus linguistics, not only tries to find examples for language trends, in this case word order

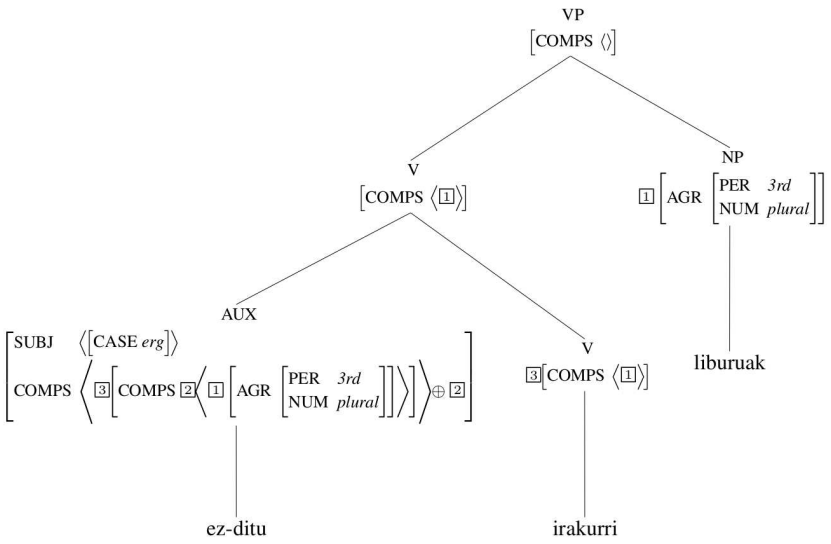


Figure 1.3: Application of the rules of a computational grammar to parse a sentence in Basque

Results of logistic regression analysis for middlefield SO/OS corpus

Coefficients:	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-3.7804	0.4803	-7.87	3.5e-15
Subject Animacy	2.5644	0.4907	5.23	1.7e-07
Object Animacy	-2.6546	0.4173	-6.36	2.0e-10
Voice	1.0646	0.5298	2.01	0.045
Perf. Auxiliary	2.0855	0.4182	4.99	6.1e-07
Subject Definiteness	1.6509	0.3942	4.19	2.8e-05
Length Difference	0.0114	0.0591	0.19	0.847

Figure 1.4: Statistical significance of several components of a language model for word order

constraints. The researchers also wished to generate a language model which, contrary to HPSG (head-driven phrase structure grammar), the grammar framework used by Crowgey and Bender (2011) and taken here as an example, does not ground on formal rules. In the latter case, the grammar can be tested automatically for completeness and is eventually tested on corpora. In the former case, the model stems from empirical information contained in the corpus. The statistical model can then be tested for significance.

Examples of statistical models in the industry are numerous, as language technology can now be found in most electronic devices: speech recognition or dictation, machine translation, spell-checking or autocompletion. The companies and research centers behind those technologies need a considerable amount of corpus data in order to build and complete their language models. The concept of “big data", as well as its impact on research practices is discussed in section 1.3.4.2 (p. 32).

Information retrieval and information extraction The most famous examples of engineers working on information extraction and retrieval systems on very large databases which are used by a very large number of users are probably those of Google and Facebook. The way linguistic information is approached in these contexts influences the global course of research

on text corpora, because the commercial success of such databases depends at least partially on the resources spent on research and development.

Additionally, since these systems are designed to be used without special knowledge by anyone, the interesting fact is that they may be used by linguists as well, to look for linguistic or numerical evidence, for instance because they make real-time text search possible and are thus close to latest language evolutions. However, the commercial interests and resulting biases of these search engines makes them unpractical for a number of reasons, which I will discuss below regarding web corpora (see p. 43).

Other disciplines Other potential corpus users are scattered among (digital) humanities in disciplines such as information sciences, communication, sociology, or history. As it falls beyond the scope of this document I will not describe their needs in detail.

One relevant usage scenario could be those studies focusing on the Web as a cultural and/or sociological phenomenon. Behavioral studies or determination of profiles among a largely unknown population of “Internet users" could highly benefit from web corpora. Recently defined disciplines such as culturomics (see p. 39) are an illustration of new research fields opening as the amount of data gathered on the Internet increases.

Conclusion: A collision of interests As a matter of fact, it can now be said that the research communities described here do not form a harmonious big picture. There is no single research field corresponding to the usage of (Web) corpora, but a multitude of research practices with diverging goals and backgrounds.

(Computational) linguistics, computer science and emerging disciplines in Digital Humanities probably form a core of highly interested potential users. In the following, particular emphasis is on linguistics and computational linguistics. However, even among this selection, goals, methodology, and faith in corpus evidence undoubtedly differ and eventually collide. These differences have a real impact on the way research data are gathered and made available.

1.2.3.2 Broad definitions of “corpus"

Consequently, the word “corpus" may have different meaning according to the scientific community employing it. Throughout this document, by “corpus" I mean text corpus, a text collection. Ostler (2008) gives a general description which can be taken as an example of a broad definition:

> “In brief, text corpora give evidence in extenso about a language, and about the content that has been expressed in that language." (Ostler, 2008, p. 458)

According to Ostler (2008), the disciplines interested in language research feature lexicography, terminology, translation, technical writing, computational linguistics, theoretical linguistics and language teaching, whereas the discipline interested in content are for instance history, literary critic, sociology, but also “advertisers and pollsters" (Ostler, 2008, p. 458).

An example of a broad definition applied to linguistics is given by McEnery (2003) as a first approach to the notion:

> “A corpus is simply described as a large body of linguistic evidence typically composed of attested language use." (McEnery, 2003, p. 449)

Baroni and Ueyama (2006) give a similar definition which seems neutral, although it is more computationally oriented, by stating that corpora are “collections of language samples produced in natural contexts and without experimental interference" (Baroni & Ueyama, 2006, p. 31).

Those definitions can be considered broad, since they include a wide range of possibilities and do not impose particular conditions on a corpus, apart from it being composed of “real", actual, attested, not experimentally biased language data.

1.2.3.3 Traditional sense of corpus and text

Definition of “corpus" according to the linguistic tradition The kind of definitions above are traditionally seen as too general, and are usually narrowed down to what is perceived as being a more linguistically sensible view, i.e. related to a more precise definition of language which includes a set of features and variants. As an example, here is the second, more precise definition given by McEnery (2003):

> “The term corpus should properly only be applied to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of a certain linguistic feature (or a set of features) via the data collected." (McEnery, 2003, p. 449)

In fact, the traditional way of seeing a corpus is that it is “not simply a collection of texts" (Biber, Conrad, & Reppen, 1998, p. 246), but at least a “carefully selected collection of texts, involving a great deal of human judgement" (Sinclair, 2008, p. 30).

The idea that “corpus compilation is to a great extent a question of judgement, taking into account the purposes the corpora are compiled for" (Johansson, 2008, p. 41) is diametrically opposed to the notion of automatically determined statistical samples on the one hand, as well as to the idea of “opportunistic" gathering of texts, described further down (see p. 19 for opportunism in a linguistic sense and p. 35 for opportunism in a statistical sense), on the other.

The traditional notion of text What’s more, the traditional notion of corpus also implies a particular perspective regarding the texts eligible to it, and thus a specific understanding of the very notion of text.

According to Atkins, Clear, and Ostler (1992), the model for the notion of text is a printed monograph or a work of narrative fiction, which sums up the main characteristics expected:

- the text is discursive and typically at least several pages long,
- it is integral,
- it is the conscious product of a unified authorial effort,
- and it is stylistically homogeneous.

Ideally, documents could be collected based on this aforementioned notion of text which makes it easy for researchers to identify, compare, or classify them.

The desiderata correspond to ideal cases that are not frequent in actual corpus uses. Even concerning traditional corpora, texts may be shorter than several pages and/or not integral, and it is unclear whether one particular “authorial effort" or class of stylistic property can

be identified. The case of traditional written corpora is described in more detail below (see section 1.2.3.3), whereas the case of web corpora, which diverge much more from these ideal conditions, is discussed further down (see section 1.4).

Classical corpus building Ideally, the texts which match the desired characteristics are eventually gathered to build a corpus.

Still according to Atkins et al. (1992), the several stages of this process can be divided as follows:

1. Specifications and design, according to “anticipated users"
2. Selection of sources
3. Obtaining copyright permissions, “a time-consuming and increasingly difficult operation"
4. Data capture (scanned by OCR or “keyboarded manually") and encoding/markup, possibly validation and error-correction
5. Corpus processing, for instance using linguistic annotation tools and a search engine

The scheme that puts design issues on the very beginning of corpus building reflects the “classical", scholarly way of building corpora. In fact, the practical obstacle consisting of obtaining the permissions to compile and use the texts comes only after the selection of texts. In that sense, it is what could be called a determinative process, opposed to what is sometimes called “opportunistic" text gathering (see p. 19) where the material conditions come first and proper corpus design resides in dealing with available sources.

On the other hand, data capture and encoding represent much more than a single step concerning web corpora. The data capture in itself is a phase that deeply impacts the nature and structure of the corpus constructed from sources on the Internet, and as such ought to require considerably more attention from the linguist’s side (see the debate on documentation p. 53).

1.2.3.4 Corpus and text typologies

General and specific/specialized corpora A single corpus is not necessarily supposed to address all the concerns and questions raised by different user communities. For this reason, many corpora are designed to a purpose, with a particular task in mind:

> “Corpora range in type from general, reference corpora designed to investigate a given language as a whole, to specialised corpora designed to answer more specific research questions." (Hunston, 2008, p. 154)

Reference corpora are general corpora that due to their construction process and structure are expected to suit most needs. The main purpose would even be for it to give a reasonable image of a language:

> “A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties

of a language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials." (Sinclair, 1996)

General corpus typology The following typology is adapted from the top-level categories of the criteria defined by Lemnitzer and Zinsmeister (2010, p. 103), which are mentioned by Perkuhn, Keibel, and Kupietz (2012) as being valuable:

- Functionality, which is described as such: the general functionality is "empirical basis for research in linguistics and/or computational linguistics" (Lemnitzer & Zinsmeister, 2010, p. 108), although one may say with McEnery (2003) that "corpora are multifunctional resources."[12]
- Language selection: mono- or multilingual
- Medium: written, oral, or mixed
- Size: size requirements have been evolving considerably since the first corpora
- Annotation level
- Persistence: monitor corpus, evolving through time, or static corpus, of fixed content
- Relation to language: reference corpus or specialized corpus (see the definition of reference corpus above)
- Availability: full text, or context through queries, due to technical restrictions or copyright issues

The distinction between synchronic and diachronic corpora evoked by Habert, Nazarenko, and Salem (1997) could be added on the top level.

The distinction evoked by Habert et al. (1997) between corpora made of full texts and corpora made of samples is not mentioned anymore in textbooks of the 2000s. In fact, it may an artifact of a given epoch. If the word "sample" is still present in corpus descriptions, it has taken a complete different meaning, mostly of a more statistical nature.

Main types of corpora Not all combinations arising from the typology above are available in practice. In fact, the number of actual corpus types found in daily practice is relatively small.

According to Renouf (2007), there were four main types of corpora from the 1960s to the 1980s and onwards:

- Standard corpora
- General and specialized corpora
- Sampled corpora
- Multi-modal, multi-dimensional corpora

[12](McEnery, 2003, p. 449)

Moreover, there are orthogonal dimensions to these types, which are linked to a special research interest, linguistic processing and annotation, and machine translation). These dimensions mostly include:

- Treebanks
- Aligned corpora
- Comparable corpora

Since this thesis mainly focuses on general and specialized corpora without any dimension added, I will not describe other corpus types any further. The notions of standard and sample are briefly tackled below (see section 1.4).

The case of variation studies and corpus taxonomy For completeness' sake, it seems necessary to mention variations studies, as a potential goal, and corpus taxonomy in terms of text categories, as a design issue.

In fact, one may add the following characteristics to the standard desiderata of digital corpora:

> "Understanding how language varies when we use it for different purposes in different situations is the central research goal of register analysis." (Biber, Conrad, & Reppen, 2006, p. 169)

Thus, general corpora in particular should include several types of language in order to enable linguists to conduct comparative studies and induce judgments about a language in general.

If several usage cases are provided, then there should be a clear classification of the texts in the corpus, corresponding to different contexts. There are corpus linguists claiming that the notion of text category is of the utmost importance:

> "Text category is the most important organising principle of most modern corpora." (O'Keeffe & McCarthy, 2010, p. 241)

Thus, it is recommended to have a clear idea of one's corpus structure and to be able to give a clear overview of it by classifying the texts, understood here as corpus components. Although the word "category" may seem obvious in some cases, such as in the difference between a newspaper article and an instruction manual, language variation "in the real world" is a manifold phenomenon which cannot usually be tackled by a single criterion.

In fact, according to Lehmann (2006, p. 23), variation is possible along the following dimensions: diastratic (speakers among different groups), diatopic (contexts), thematic, communication type and text genre, media type.

Regarding web corpora specifically, the communication and media type do not vary significantly. It is possible to record a number of cases and study them all throughout the corpus. The diatopic variation is already more difficult to capture efficiently: since the Web is evolving rapidly and since new trends in communication on the Internet also appear at a fast pace, not all contexts are known in advance, because there are no such things as finite, strictly definable contexts.

The diastractic and thematic dimensions are particularly difficult to follow, since there are probably many different speaker groups the researchers do not know about, let alone those who are constantly appearing and evolving. The depending thematic information could be useful for all purposes but it is difficult to aggregate based on so many unknowns.

1.3 Corpus design history: From copy typists to web data (1959-today)

1.3.1 Possible periodizations

The notion of “linguistic data" is relatively new compared to the time-honored philological tradition. It most probably dates back to the beginning of the 20th century, and as such it can be considered too young to be completely developed (Lehmann, 2006, p. 13).

The history of machine-readable corpora is even shorter, most notably because the very first computers were not primarily designed to process text:

> “In its modern, computerized form, the corpus has only existed since the last 1940s." (McEnery, 2003, p. 452)

Tognini Bonelli (2010) sees four main generations in electronic corpus building. First, the beginning, when texts were transliterated. Second, a change around 1980, opening the “decade of the scanner" and paving the way (third) to a “First Serendipity" after 1990, with text available as a “byproduct of computer typesetting". Last, the so-called “Second Serendipity" beginning with the new millenium, when “text that never had existence as hard copies becomes available in unlimited quantities from the internet".

This description is a bit schematic, mostly because of the rounded dates and the fact that the history only begins in 1960. However, the technological evolutions and the changes in access to text it retraces are acknowledged. There is a broad consensus about the fact that corpus design saw a major change around 1990, when bigger corpora were made available, mostly because of technological advances:

> “During the 1980s, the number of corpora available steadily grew as did the size of those corpora. This trend became clear in the 1990s, with corpora such as the British National Corpus and the Bank of English reaching vast sizes." (McEnery, 2003, p. 452)

Another change lies in a methodological shift towards what can be called “linguistic technologies", which also happened at the end of the 1980s. Contrary to the advances described above, it has nothing to do with technology in itself, but rather with communication of research objects and interests between communities. Other potential users, who would later be named computational linguists or members of the natural language processing community, were interested in using corpora as a testbed, while at the same time corpus builders realized how useful technologies from this field could be to corpora.

> “There was an increasing interest in the use of corpora for computer-based linguistic technologies, though the main developments were to come later." (Johansson, 2008, p. 49)

Besides, it is a sure fact that the amount of available text for linguistic analysis has become larger and larger, so that Leech (2006) speaks of linguists as “inhabiting an expanding universe". This universe has been expanding rapidly since the emergence of web as potential corpus contents. The shift took place around the year 2000, as I detail below (see p. 44).

1.3.2 Copy typists, first electronic corpora and establishment of a scientific methodology (1959-1980s)

1.3.2.1 Influential corpora

An influential pre-electronic corpus: the SEU (1959) Compilation of the Survey of English Usage (SEU) corpus began in 1959 at the Survey of English Usage (University College London) under the direction of Randolph Quirk (Meyer, 2008, p. 10). Even though data collected on pieces of cardboard is radically different from “modern" electronic corpora, for instance because it is much more difficult to handle and to transform, the corpus can be considered to have been influential due to its scope and due to the design decisions and later work of its compilers:

> “It was the first corpus created expressly for use by those other than its creators. In addition, the principles of corpus creation that guided its creation are still very relevant and actively applied to those building modern corpora." (Meyer, 2008, p. 12)

There is real foresight in the statement by Quirk on former grammars, on the one hand because of the focus on empiricism and not introspection, and on the other hand because of the assumed intent of finding norms and not necessarily rules:

> “Their generally eclectic use of written source materials too often leaves unclear the distinction between normal and relatively abnormal structures and the conditions for selecting the latter." (quoted by Meyer (2008, p. 10))

What legitimizes a structure is not a decision by a grammarian, it is its regularity in general language use. In addition, Quirk’s position on regularity is not a binary decision, the constructions which stay out of the grammar are not believed to be impossible or false, they are “relatively abnormal".

In order to get a general view on language and norms, the SEU corpus comprised a range of different types of written and spoken texts (Meyer, 2008, p. 11) amounting to a total of one million words, divided into 5000 word samples (Meyer, 2008, p. 12). Due to material limitations, text excerpts were seen as a way to include more text types:

> “Individual samples consist mainly of text excerpts. Sampling texts in this manner ensured that many different examples of a given text-type could be included as well as a range of different speakers and writers." (Meyer, 2008, p. 12)

The compilation of a corpus in order to establish a reference for the study of language is an idea which has been productive in the decades following the SEU.

Selection of texts and influence of the Brown corpus (1960s) The Brown corpus belongs to what may be called the "Brown family" (Xiao, 2008), since roughly comparable corpora were built along the same lines (most notably for British English, Indian English, Australian English, and New Zealand English). It is a successful model, as it was constructed with comparative studies in mind (Xiao, 2008, p. 395).

The Brown corpus deals with written American English and contains 500 samples of approximately 2,000 words. These samples are spread among 15 text categories, which were defined and chosen before corpus construction. In sum, the corpus is the result of design decisions and random sampling methods:

> "The selection of texts was based on a combination of considered judgement (the choice of text categories and their weighting) and random sampling." (Johansson, 2008, p. 36)

The corpus is synchronic and it is supposed to give a representative/balanced image of English:

> "The corpora of the Brown family are balanced corpora representing a static snapshot of a language or language variety in a certain period." (Xiao, 2008, p. 396)

According to Johansson (2008), the significance of the Brown corpus is threefold. First, it "established a pattern for the use of electronic corpora in linguistics", since it is one of the first of its kind. Second, the sampling method and its documentation allowed for fruitful subsequent corpus research.[13] Last, the fact that the corpus was made available to others and that corpus research was conducted beyond a single research institution also had a exemplary value.

1.3.2.2 The establishment of corpora as a scientific object: Decisive characteristics of "traditional" reference corpora

By "traditional" reference corpora I mean general text collections fulfilling the criteria of traditional text and corpus definitions (see above p. 11), which are established as valuable and linguistically defensible corpora, e.g. the Brown corpus or the BNC in English linguistics.

Motivation and objectives Among the very first steps in order to create a corpus, the need to establish that the corpus is a valid scientific object and that it responds to motivated research needs seems to be paramount:

> "The first issue for the Birmingham Corpus, as for the earlier 'standard' corpora, was theoretical: how to create a body of text which could be claimed to be an authoritative object of study." (Renouf, 2007, p. 32)

In fact, the notion of objective science can have a strong influence on the way research is perceived and evaluated (Lehmann, 2006, p. 10).

This "authoritative" and ideally indisputable character is due to the fact that at the age of "classical" corpora the texts are approached according to a "corpus-based" methodology, in

[13] Johansson highlights the "care which was taken to systematically sample texts for the corpus and provide detailed documentation in the accompanying manual" (Johansson, 2008, p. 38)

order to provide linguistic evidence. As a consequence, the fact that examples or statistics extracted from the corpus are scientifically valid has to be above any doubt.

This necessitates the justification of corpora, simple size or diversity are not sufficient in order to be taken as scientific evidence:

> "For most linguists, then, a corpus cannot be equated with just a large collection of texts or citations, but needs to be justified in linguistic terms." (Tognini-Bonelli, 2001, p. 55)

In the terms of the 20th century this justification is an example of the blur between a corpus and its use, and the difference in corpus tradition may explain part of the incomprehension between the approaches mentioned above. The definition of a corpus also includes restrictions in terms of text availability and processing power, as the inclusion of whole text or text samples for instance is disputed:

> "We can see that they all agree that a corpus is a collection of language text, though not necessarily texts. Aarts talks of samples, and EAGLES of pieces of language. Francis alone talks of texts, and Atkins et al. appear also to see a corpus as restricted to a collection of whole texts (though most corpora, including Francis' Brown Corpus, are not)." (Tognini-Bonelli, 2001, p. 53)

Due to the restrictions and the need for justification mentioned above, and since a corpus cannot pretend to include all the existing texts and/or to encompass all existing varieties, the notion of corpus as a valid sample of the target population, variety or genre naturally arises. For Tognini-Bonelli (2001), the debate on validity leads to three different issues to be tackled: authenticity, representativeness and sampling.

Sampling The sampling issue is linked to the need for "authoritative" corpus building described above:

> "A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language." (Sinclair, 1996)

Among others, the Brown Corpus and the British National Corpus used text samples for pragmatic reasons, in order to restrain the size of the whole corpus and to make it easier to process, or because of copyright restrictions. Concerning the written texts of the British National Corpus, each text is limited to 40,000 words or 90% of the total content, whichever is reached first.

Copyright restrictions may still apply and impede the use of whole texts, but the sampling due to processing speed limitations is no longer required.[14] Nowadays, 2000-word samples are seen as an untenable limitation, but the idea that all text samples should be of the same size is still present. Since the beginning, samples have been seen as the lesser evil, a way to obtain enough variety at the cost of fragments and the loss of complete texts (Sinclair, 2008, p. 25).

[14] "The issue of the number of words that constitutes a proper sample of text is one that, happily, is now a matter of history." (Sinclair, 2008, p. 24)

Background on representativeness/representativity The idea of representativeness in turn is linked to the sampling processes as well as to the corpus being considered a general view on language.

First, a corpus can be seen as a necessarily limited amount of text or utterances compared to everything that is being said and written. In this perspective, any given corpus is a sample in itself, whether it is composed of samples or not. Second, questions may rise regarding the general value of studies performed on that corpus. Strictly speaking, a corpus which is not representative only speaks for itself, and precludes a potential generalization of results to language as a whole. That is why representativeness may be seen as a useful and/or paramount corpus feature.

Leech (2006) gives an operational definition of representativeness/representativity as a goal:

> "'Representative' means that the study of a corpus (or combination of corpora) can stand proxy for the study of some entire language or variety of a language. It means that anyone carrying out a principled study on a representative corpus (regarded as a sample of a larger population, its textual universe) can extrapolate from the corpus to the whole universe of language use of which the corpus is a representative sample." (Leech, 2006, p. 135)

A priori and a posteriori representativeness, traditional and opportunistic perspectives on corpora Biber (1993) adopts an a priori perspective and makes an issue of corpus design of representativeness, whereas Sinclair generally opts for an a posteriori view on the topic; for him, the issue of representativeness can be tackled once the corpus has been built. In fact, Sinclair advises to use corpora despite their limitations:

> "Since language text is a population without limit, and a corpus is necessarily finite at some point; a corpus, not matter how big, is not guaranteed to exemplify all the patterns of the language in roughly their normal proportions. But since there is no known alternative for finding them all, we use corpora in full awareness of their possible shortcomings." (Sinclair, 2008, p. 30)

The willingness to use corpora despite potential caveats such as the fact that they may not be representative can be characterized not only as an a posteriori perspective but also as an opportunistic view on corpora.

Practical response to the need for representativeness Most researchers agree on the fact that although representativeness is part of the scientific justification strategy, it is an unattainable goal in practice:

> "Through the 1980s and early 1990s, though it was generally accepted among corpus creators that representativeness was unattainable, it was felt necessary to present selectional criteria in those terms." (Renouf, 2007, p. 32)

> "Representative and balanced corpora are a theoretical ideal corpus compilers constantly bear in mind, but the ultimate and exact way of compiling a truly representative and balanced corpus has eluded us so far." (Gries, 2009, p. 1232)

According to Hunston (2008, p. 162), there are three possible responses to the problems posed by the notion of representativeness. The first one is to avoid the idea, to consider the corpus as a collection of different registers, without claiming comprehensive coverage. The second one is to "acknowledge the problems but do the best that is possible under the circumstances and be transparent about how the corpus has been designed and what is in it". A solution would be to let the corpus users assess the degree of representativeness or to take advantage of an hermeneutic cycle (see p. 24). The third possibility consists of attempting to include texts from as many different sources as possible into the corpus and treating the result as a collection of sub-corpora rather than as a single entity.

Balance/balancedness The notion of balance is closely linked to the notion of representativeness:

> "This 'balanced' quality has frequently been claimed for corpora such as the Brown Corpus or the British National Corpus or ICE-GB, which have been carefully designed to provide sufficient samples of a wide and 'representative' range of text types." (Leech, 2006, p. 136)

In fact, balance impacts the internal composition of a corpus, it may be seen as a decision between the number of texts and the number of tokens to be included (Hunston, 2008).

It is especially useful as it allows for (synchronic or diachronic) comparative studies:

> "The real benefit of a balanced corpus is that each of its various components is large enough to make comparisons feasible." (Hunston, 2008, p. 164)

However, it does not seem to be strictly necessary for the establishment of a corpus, the LOB corpus for instance makes no strict claims of representativeness (Johansson, 2008, p. 40).

Moreover, the more lenient notion of "exemplary corpus" as defined by Bungarten (1979) can be seen as a remedy of the constraints of representativeness, where the consensus among specialists for instance accounts for the validity of corpus design.

Authenticity Once again, newspaper articles or narrative texts for instance are a paragon for the notion of texts (as defined p. 11), so that they may be used as a reference for texts to include in a corpus. Concerning such texts, one can indeed claim that "the default value for quality is authentic" (Sinclair, 1996), because they were produced in what Gries (2009) calls a "natural setting":

> "For example, many corpora consist to a large degree of newspaper articles. These are of course often included for convenience's sake, but they also meet the criterion of having been produced in a natural setting because journalists write the article to be published in newspapers and magazines and to communicate something to their readers, but not because they want to fill a linguist's corpus." (Gries, 2009, p. 1232)

Thus, authenticity is a classical criterion which texts included in corpora have to satisfy. It addresses the very production conditions of utterances, for instance to what extent a conversation has to be "non-fictitious" and to have "really" existed:

> “Meanwhile, whilst it was clear to most corpus creators from the outset that dramatic dialogue did not constitute an authentic use of speech, and was thus to be excluded, the issue of how to classify conversation within novels has remained largely unresolved." (Renouf, 2007, p. 46)

In sum, authenticity is a characteristic that is wished for, but which is not equally distributed among text genres or even within one particular genre. Articles published in magazines are most probably authentic, whereas dialogs in fiction are much more difficult to classify.

1.3.2.3 Summary

There is an intention behind a corpus, regarding the object of interest as well as concerning the establishment of the corpus as a valuable and scientifically defensible resource.

Beside a full-fledged typology, there are two main types of corpora: reference corpora and specific corpora. Reference corpora are supposed to be big and diverse enough to allow for fruitful studies and avoid the biases mentioned by Biber (1993): random error (sample too restricted in size), and bias error (systematic differences in regard to target population).

Corpus building includes a number of steps. Text sampling used to be a constitutive part of text corpora, most notably for practical reasons, but not anymore. Text selection has always been an important step, for which a balance has to be found in terms of content size and diversity.

Although time has passed, reference corpora builders still have to deal with similar issues, and they still use comparable methodology. Corpus representativeness, for instance, has always been considered to be crucial but is still problematic in practice. A given corpus is a sample of language, whose global stratification is unknown, and part of its justification resides in the possibility of extrapolating language facts found within the corpus to general assumptions concerning language use.

1.3.3 Digitized text and emergence of corpus linguistics, between technological evolutions and cross-breeding with NLP (from the late 1980s onwards)

1.3.3.1 The empirical trend of corpus linguistics

There are different reasons for and different responses to the “empirical trend" (Sampson, 2013) in linguistics that itself is behind a discipline or research current called “corpus linguistics". What it is and to what extent it forms a coherent subtype of linguistics remains unclear. A sure fact is that despite the rise of empirical methods “corpus linguists" are nowadays still a minority among linguists:

> “The recent rise in interest in corpus-based research methods has been caused in part at least by a reaction against that unempirical style of linguistic research. But linguists who would identify themselves as ‘corpus linguists’ are still, surely, a minority, and it is not clear to what extent the discipline as a whole has genuinely reformed and accepted the logic of empirical methodology (normative empiricism)." (Sampson, 2013, p. 284)

Thus, the empirical trend in general linguistics only concerns a minority of researchers at its full extent, and probably more researchers as a background tendency towards attested data. This statement needs to be completed by two pieces of information: first, the trend is not completely new, it corresponds to a research current which was dominant in the 1950s; and second, empiricism only resurfaced in the 1990s, which means it is less than a generation old:

> "The 1990s have witnessed a resurgence of interest in 1950s-style empirical and statistical methods of language analysis. Empiricism was at its peak in the 1950s, dominating a broad set of fields ranging from psychology (behaviorism) to electrical engineering (information theory). At that time, it was common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words." (Church & Mercer, 1993, p. 1)

Church and Mercer (1993) are strong advocates of empirical approaches and data quantity. There is an antagonistic positioning in their very definition of the discipline, in their rationale in favor of co-occurrences, for instance, as well as in the designation of an "unempirical style" and of a "reaction" by Sampson (2013). This is discussed below in section 1.3.3.4. This positioning is not trivial, because it is closely tied to the evolution of technology and methodology.

1.3.3.2 The role of technology, experimental traces, and instruments

There have indeed been developments that changed the way linguistics and computational linguistics are considered. These took place both on a technological and on a methodological/paradigmatic level, the first increasingly impacting the latter. The trend which started with natively digital text in the 1980s has gained influence ever since (cf p. 31 for more recent developments).

An influential technological evolution At first, computers were only considered a tool, but in the course of technological evolution they began to affect the methodological frame of inquiry (Tognini Bonelli, 2010, p. 17). In fact, computerized corpora actually shaped corpus linguistics as a discipline, and not only from a technical point of view:

> "It was the revolution in hardware and software in the 1980s and 1990s which really allowed corpus linguistics as we know to emerge." (McCarthy & O'Keeffe, 2010, p. 5)

The trend towards computerized or computer-assisted science is not unique to linguistics, a wide range of disciplines are touched by this evolution:

> "Informatics need not merely support traditional ways of conducting inquiry in a particular discipline, but can fundamentally change the development of a discipline." (The Royal Society Science Policy Center, 2012, p. 31)

The increasing availability of research traces as a dispositive toward scientific productivity Berthelot (2003) defines the scientific process as a sedimentation of notes and data.

He sees a continuous flow from experiment to abstraction, where experiment protocols and (partial) records progressively transform into synthetic, anonymous, and stabilized results.[15]

Latour (1985) also describes the evolution of science in the 1980s as a flow operating on research information and enabling researchers to deconstruct, transform, and synthesize it. He sees a particular value in the changing of scale, the combination of traces, the incorporation of inscriptions in a text, and the fusion with mathematics.[16]

Thus, there is a particular context of evolving scientific paradigms and growing importance of the transformation, combination, and exchange of information gained during experiments since the 1980s. The systems outlined by Berthelot and Latour explain the growing popularity of an empirical trend which produces traces that may in turn be processed in a computational fashion. The cross-breeding with mathematics evoked by Latour (1985), and, in the case of linguistics more precisely the impact of computer science and NLP, also sheds a new light on what can nowadays be called research data. The data paradigm and the further evolution of sciences into technosciences are described more in detail in the next section (see p. 31).

In concrete terms: instrumentalized linguistics For now, let us focus on the notion of experimental traces. Annotation and export formats, for example the markup language XML, allow for greater accessibility of research objects, and not merely a research report. Multiple judgments can for instance be literally incorporated into a text, similarly to the process described by Latour. Text enrichment through annotation, the adding of information, is precisely the kind of activity that produces a great number of traces which have to be made available, to humans but also to machines when possible. The text then becomes an experimental reality and manipulable proof substrate.[17]

It is not only about designing an ordered data archival system; the process described above is part of the establishment of a dispositive toward scientific productivity. In that sense, the experimental "field" eventually includes its dispositive, residing in metrics, instruments, and reading results. It ceases to be a "field" to become an organized restitution allowing for the testing of hypotheses and providing background for reasoning.

Even better, the field as seen through such instruments can offer supplementary dimensions of analysis, so that the couple dispositive/field can become much more productive than separate approaches. Mathieu Valette for instance evokes the advantages of an "instrumentalized text science".[18] This science can allow for the construction of new observables which would have been invisible otherwise.[19]

[15] "[Un] processus historique par lequel une multitude de comptes rendus, datés et signés, d'expériences et de formules partielles se transforme progressivement en résultats synthétiques anonymes et stabilisés." (Berthelot, 2003, p. 29)

[16] "Mobiliser, fixer immuablement les formes, aplatir, varier l'échelle, recombiner et superposer les traces, incorporer l'inscription dans un texte, fusionner avec les mathématiques." (Latour, 1985)

[17] "Il faut donner aux phénomènes une forme qui soit telle que l'on puisse, en la retravaillant, gagner sur eux plus d'informations qu'on y a mis." (Latour, 1985)

[18] (Valette, 2008, p. 11)

[19] "Les grandes masses de données textuelles ou documentaires nécessitent, pour être analysées et décrites, des dispositifs expérimentaux et des instruments *ad hoc*. Cette instrumentation permet de construire de nouveaux observables qui seraient demeurés invisibles autrement." (Valette, 2008, p. 11)

1.3.3.3 Corpus-based, corpus-driven and cyclic approaches

Corpus-based and corpus-driven As linguistics becomes instrumentalized, and since new observables are constantly being created, there is an epistemological uncertainty concerning the way corpora are to be approached as well as concerning the impact they have on an hypothesis. The question is whether corpora are to be used in order to find evidence for known phenomena. This in turn raises the issue of attested data and for instance “good" examples in lexicography. In short, one may wonder whether corpora are strictly destined to perform hypothesis testing, or whether something is to be learned from the corpus, by a so-called data-centered approach. While both approaches could be used alternatively or one after another, they often exemplify a gap between two different views on language and research processes.

The expressions of “corpus-based" and “corpus-driven" research are detailed by Tognini-Bonelli (2001) in order to give an account of the diverging methodologies.

Durand (2009) uses the metaphor of the telescope to state that some linguists use corpora to observe a given language state, while others don’t use them to validate hypotheses in the first place, but as the starting ground for discovery:

> “It is often said that corpora are the equivalent of the telescope in the history of astronomy. Many of the hypotheses concerning the nature of the universe were in place before telescopes were invented. But progressively the observations they made possible proved crucial in the validation or invalidation of various theories. However, it should also be observed that a number of linguists have taken a radical route concerning the nature of data in linguistics." (Durand, 2009, p. 13)

The use of telescopes that Durand (2009) favors would be what corpus-based linguistics is supposed to mean: there is a pre-existing theory. In fact, the problem is more complicated than that, since technology in the case of a telescope is nothing without proper observation protocols. What’s more, there is so much to see in a night sky that the tendency is for theories to be confirmed or rejected because astronomers usually need something specific to look for.

Cyclic corpus revision Other researchers are in favor of a cyclic model where corpus data provide a feedback for hypotheses but only to amend them or generate new ones, in what could be called a “hybrid" approach:

> “’Soft’ linguistic research on text corpora may be characterised as three distinct stages, what we call the ‘3A perspective’ (Annotation – Abstraction – Analysis). Each of these stages represent a potential source of error between the original text and the evaluation of hypotheses. Each is knowledge intensive and approximate, and therefore cyclic." (Wallis & Nelson, 2001, p.311)

In a similar approach, Pincemin (2006) speaks of an “hermeneutic circle" of interpretation to clarify that interpretation has an impact on analysis, and that seeing the results of any given stage constitues feedback that may allow for adjustments in coding and processing methodology.[20]

[20] “L’interprétation est présente à toutes les étapes du travail sur corpus. Interprétation a priori au moment de la constitution du corpus, et avec la conception des opérations d’analyse à pratiquer ; interprétation a posteriori pour l’exploitation des résultats produits. Mais la pratique interprétative procède par retours et ajustements, elle n’échappe pas au cercle herméneutique : ainsi, la lecture des résultats motive très naturellement une reprise du codage et une réorientation des traitements." (Pincemin, 2006)

Atkins et al. (1992) mention what they call a “hermeneutic cycle" in the case of representativeness and correct sampling, which for them cannot be achieved on a theoretical basis, i.e. before the corpus has been built. Cyclic repetitions of sampling processes are needed to adjust the corpus over and over again:

> “In the light of this experience and feedback the corpus is enhanced by the addition or deletion of material and the cycle is repeated continually." (Atkins et al., 1992, p. 5)

The same is true for balance, which depends on feedback from users.[21]

1.3.3.4 Corpus linguistics seen as an artifact of academic discourse

Studies based on large corpora have led to new paradigms and in fact to corpus linguistics, which certain scientists are willing to consider an autonomous entity. However, this autonomy is not obvious.

Arguably, the expression “corpus linguistics" was first used in English by Jan Aarts in 1983, and in Dutch a few years before (Johansson, 2008, p. 34). There is definitely an international aspect to the development of corpus linguistics, as well as a central role of the English language and English corpora (Johansson, 2008, p. 35).

A productively antagonistic, but unclear positioning According to Léon (2005), there is a discourse regarding the theoretical foundations of corpus linguistics presenting this research field as systematically opposed to the ideas of generative grammar starting early on, and even more so in the 1990s, when corpus linguistics began to be institutionally recognized. According to this discourse, corpus linguistics is supposed to follow other scientific interests and to introduce a fourfold shift of focus:

- Focus on linguistic performance, rather than competence
- Focus on (concrete) linguistic description rather than on (abstract) linguistic universals
- Focus on quantitative, as well as qualitative models of language
- Focus on a more empirical, rather than a rationalist view of scientific inquiry.

In the early 1990s, the academic discourse was aimed at a legitimization of corpus linguistics seen as an autonomous discipline by supporting a retrospective construction of the research field[22]. To do so, pioneer works in computerized corpora were omitted, for instance corpora in machine translation or dictionary making[23].

[21] “Controlling the ‘balance’ of a corpus is something which may be undertaken only after the corpus (or at least an initial provisional corpus) has been built; it depends on feedback from the corpus users, who as they study the data will come to appreciate the strengths of the corpus and be aware of its specific weaknesses." (Atkins et al., 1992, p. 14)

[22] (Léon, 2005, p. 46)

[23] *ibid.*, p. 47

> “Two retrospective constructions were forged at the moment when NLP was technologically ready to invest in the field of corpora: a theoretical anti-precursor, one of the most famous theoretical linguist, i.e. Chomsky; and a technical precursor, in fact a product, the Brown corpus." (Léon, 2005, p. 47)

This oppositional positioning of corpus linguistics is also described by Teubert (2010), for whom corpora were first used in so-called “applied linguistics", more specifically in dictionary making and language teaching:

> “Corpus linguistics was originally the response to the need of teaching English as a foreign language. Traditional dictionaries, with their focus on the single word in isolation, could not tell its users how to use a word. Corpus linguistics was an approach that could remedy this deficiency. It was not conceived as a reaction to the self-defeating endeavours of Chomskyan and cognitive linguistics to describe the way in which the individual mind processes language. It was firmly situated in what is called in the British context applied linguistics, with its focus on language teaching and dictionary making." (Teubert, 2010)

Accordingly, the purpose was to find salient and/or self-explaining uses of words or grammatical rules to provide a better understanding of how language works and thereby make it more easily available to others, for instance language learners.

If the approach, practices and technologies used in corpus linguistics seem to justify the creation of a particular discipline, the question remains open whether it is an experimental discipline per se and not a subfield or research current of linguistics. Hence Mathieu Valette concludes that corpus linguistics will most probably never become established as an academic discipline. [24] He rather thinks that linguistics as a whole must take a stand on the latest developments engendered by technological changes.

Marcel Cori sees the problem in terms of epistemological characterization of language and research in linguistics. He considers corpus linguistics to be a double characterization because it defines itself both as research on language and as applied linguistics (Cori, 2008, p. 105). In short, practice gives a feedback to theory and pretends to change its extension.

Even if it is a construct, this opposition has been productive in the way that it has helped to set the ground for another way of seeing acceptability judgments on the one hand and finding and proving linguistic theories on the other. Nonetheless, it may explain both the broad consensus among the community concerning the proper definition of corpus linguistics and the open debate concerning its scope, whether corpus linguistics is a discipline per se or rather a methodology:

> “The main area for debate relates to the scope of corpus linguistics, with some researchers arguing that it is more than just a methodology and instead should be considered a new branch of linguistics" (Anthony, 2013, p.142).

In addition to the “expanding universe" linguists live in (see p. 16), the corpus linguistics community is rapidly expanding, within its countries of origin as well as geographically: “corpus linguistics, although it largely derives from Western Europe and North America, is no longer restricted to those areas" (Ostler, 2008, p. 482).

[24] “La linguistique de corpus ne sera, selon toute vraisemblance, jamais établie en discipline académique." (Valette, 2008, p. 9)

1.3.3.5 Example 1: the BNC (mid 1990s)

Intention The scope of the British National Corpus (BNC) is very wide, as it is supposed to include "as wide a range of modern British English as possible" in order to "say something about language in general" (Xiao, 2008, p. 384). In that sense, it is a reference corpus as well as a general corpus, exemplifying that in the BNC's approach no productive distinction is made between reference and general corpora.

The emphasis lies rather on properly acknowledged and classified language varieties, which are taken into account with two objectives. First of all, the corpus is expected to be exhaustive ("[the] corpus could be regarded as a microcosm of current British English in its entirety") and to allow for cross-type comparisons (Burnard (ed.), 2007)

Thus, the corpus is seen as a "microcosm", the miniature replica of real world trends and phenomena, which were made available and exploitable by a proper corpus construction methodology and rich metadata.

Composition The BNC features approximately 100 million words. It is mainly composed of written texts (90%), but also features transcripts of speech (10%). The diversity of the corpus is hard-coded, it does not evolve in the course of time to account for new text types for instance.

Concerning written texts, the categories are established following three criteria: "domain" (the content type), "time" (the date of production) and "medium" (the type of text publication) (Xiao, 2008, p. 384). If applicable, metadata also include information about the conditions of the production of the texts, for example through the notions of "authorship" and "target", which have later been adapted to web corpora by Sharoff (2004).

Most of the time there are no full texts, just extracts. The texts are sampled so that "no extract included in the corpus exceeds 45,000 words" (Burnard (ed.), 2007). The decision was made mostly due to copyright reasons but also because of balancing. A curious fact, which highlights the difficulty to provide large text collections with detailed metadata, is that the "sampling type" of texts, i.e. whether the beginning, the middle, the end of a given text was selected or whether the extract in the corpus is a composite part, is unknown for nearly 40% of the texts (Burnard (ed.), 2007).

Additionally, for most dimensions such as the "authorship" mentioned above, the metadata fields are left empty because they cannot be filled with certainty. Thus, a large part of the texts in the corpus come with "basic" metadata, the rest being unknown, the same as in web corpora.

A corpus for all? However, even a general corpus of this kind cannot suit all needs, so that the "general" purpose does not necessarily include all imaginable research fields, despite the intention of the corpus creators to broaden the user spectrum:

> "The British National Corpus (BNC) is considered to be a general-purpose, general-language corpus, designed to be used for NLP as well as grammar studies or as the basis for large monolingual dictionaries. However, one would hardly use the BNC for conducting studies of language acquisition or for diachronic investigation." (Geyken, 2007, p. 25)

No corpus construction cannot accomodate all user types, which can be compensated for by a generous interface designed with several research scenarios in mind. In that sense, the

fact that the BNC is accessible online via a web interface (Xiao, 2008, p. 385) is defining, since it enables users to work on one particular subset of the whole.

1.3.3.6 Example 2: the DWDS core (mid 2000s)

Introduction The DWDS core corpus (*Kernkorpus*) is part of the DWDS project[25], whose main purpose is "to serve as the empirical basis of a large monolingual dictionary of the 20th/21st century" (Geyken, 2007, p. 25). In order to reach this goal, representativeness is a desideratum. However, without mentioning the theoretical debate on this notion, there are practical difficulties (mainly the small amount of digitized texts) which make it necessary to lower expectations and turn to the "more modest" notion of "balance".[26]

There are in sum three criteria which the DWDS core corpus is expected to satisfy: balance, size (large enough), and it "must also contain a considerable amount of influential and important literature" (Geyken, 2007, p. 26).

Structure A total of five genres was identified, "guided by the practical consideration that fewer genre distinctions make the daily corpus work easier" (Geyken, 2007, p. 28): journalism (27% of the corpus), literary texts (26%), scientific literature (22%), other nonfiction (20%), and transcripts of spoken language (5%).

Corpus building followed four main steps, where the greatest problem was the "administrative task of negotiation copyright clearance" (Geyken, 2007, p. 28):

- text selection
- copyright acknowledgments
- digitization and conversion to a structured format
- text sampling

In fact, in accordance with the steps described by Lemnitzer and Zinsmeister (2010), the texts were selected without verifying that they were available [27], using an approach which is typical for reference corpora. The texts were selected by different experts according to their respective genre: researchers for science, specialists of German literature and official lists (books acknowledged for their literary quality), lexicographers.

Different strategies were designed in order to deal with copyright restrictions (Geyken, 2007, p. 30). First of all, text samples are used instead of integral texts if needed. Second, the querying interface allows for variable context windows, from seven words to one or three sentences, or even a paragraph depending on the agreement. Third, copyrighted texts are only accessible to logged-in users, who agreed to the non-commercial use of text material. Last, named entities can be anonymized, which can be important for private letters for example. As a consequence of these measures, the project obtained sufficient permissions from fifteen different publishing houses to make about 71 % of the corpus available, whereas 29% are for internal use only.

[25] `http://www.dwds.de`

[26] "There are also practical obstacles: too many German texts of the 20th century are not yet digitized; the situation is even worse for the spoken language. Therefore, many corpus linguists have abandoned the notion of representativeness and replaced it by the more modest notion of 'balance'." (Geyken, 2007, p. 25)

[27] "The text selection was conducted independently of the copyright status." (Geyken, 2007, p. 30)

At the end of 2005, the total text base for the DWDS Kerncorpus comprised 272,215 documents with 254,293,835 tokens (Geyken, 2007, p. 34). From this text base a balanced text corpus of about 100,000,000 tokens is extracted with a sampling procedure, using an 'initial corpus' comprising texts of high interest as well as a mean value per decade and an acceptable deviation. The corpus is annotated in order to provide morphological information, lemmatized forms, part of speech tags, as well as syntactic information.

1.3.3.7 The theoretical and technological blur affecting corpora

Corpus goals As a consequence of the various user profiles and scientific traditions involved, there are certain theoretical and technical blurs which explain why the very essence of a corpus can be put into question and how corpus construction methods can be challenged.

The first concerns the notion of corpus and its afferent goals. The notion of "good" corpus is questionable, because it first has to be a corpus, with describable intentions, design and characteristics:

> "McEnery and Wilson (following others before them) mix the question 'What is a corpus?' with 'What is a good corpus (for certain kinds of linguistic study)?' muddying the simple question 'Is corpus x good for task y?' with the semantic question 'Is x a corpus at all?'" (Kilgarriff & Grefenstette, 2003, p. 334)

Furthermore, even if a corpus has been designed for a particular use, the original goals or at least the original setting can change in the course of the hermeneutic cycle described above p. 24, so that there definitely are multi-purpose corpora, intentionally or not.

The corpus can even be seen as an artifact of particular goals:

> "[The] arguments that a particular corpus is representative, or balanced, are inevitably circular, in that the categories we are invited to observe are artifacts of the design procedure." (Hunston, 2008, p. 164)

The corpus as a blend of research goals and processing issues The second blur consists of a certain blindness towards the tools of analysis. Tools may give an apparent immediate access to corpora, which let them become invisible to the user and let them forget that a given set of corpus processing and querying tools can alter the original material significantly. In the end, it is as if it was possible to look through a corpus without their mediation:

> "There is a continuing tendency within the field to ignore the tools of analysis and to consider the corpus data itself as an unchanging 'tool' that we use to directly observe new phenomenon in language." (Anthony, 2013, p. 145)

According to Anthony (2013), the first reason for the unsuspected value of processing tools is the heterogeneity of corpora and the strong variability of content quality in itself, without even taking linguistic processing into account:

> "One reason for blurring the separation between the data and tools of corpus linguistics is that the data itself can vary tremendously in quality and quantity depending on the research design." (Anthony, 2013, p. 144)

A second reason lies in the fact that tools can be seen as black boxes to the user or at least as a dense intermingling of research goals and natural language processing technicalities:

> "Another reason is that the tools used in corpus linguistics are software based, and thus, abstract in nature." (Anthony, 2013, p. 144)

Thus, it is notably difficult for the end user to get an overview of what happens during a particular processing stage, all the more since the software accessible to the end user potentially differs in its nature and function from the tools used by corpus compilers, for instance a linguistic search engine on the one hand, and tools giving feedback about corpus composition and processing efficiency on the other hand.

One may add that the designers of tools themselves try to keep the engineering effort in the background in order to be able to reach a larger spectrum of users. The need for a somewhat simple and direct user experience creates the illusion that there is a direct and transparent access to the corpus. In a analogous way, Auroux (1998) remarks that language itself does not seem to need particular tools to observe and manipulate it.

Moreover, it most frequently requires direct implication in corpus building projects to appreciate the impact of tools in general and linguistic preprocessing and processing in particular:

> "Without a deep knowledge of these different aspects of software design and their impact on data analyses, it is easy to forget the crucial role that tools play." (Anthony, 2013, p. 144)

However, this knowledge cannot be expected from each and every user.

The methodological openness of computational linguists regarding different theoretical frameworks is sometimes also seen as an opportunism lending itself to criticism. Cori (2008) for instance see the integrative and applicative character as an advantage from an engineering point of view, but a clear downside from a scientific research perspective[28]. Cori (2008) explain the haziness on the epistemological side precisely by the blend of processing issues and research methodology[29] which is at the core of corpus-based studies, so that one can say that this limitation of corpus linguistics is noticeable among corpus enthusiasts (Anthony, 2013) as well as among critics (Cori, 2008).

A third uncertainty seems to mingle methodology and data characteristics, i.e. the so-called "qualitative" or "quantitative" methods, which according to Loiseau (2008) do not oppose proper methods in a scientific way but main data types in a technical sense[30]. This is also a consequence of a blend of research goals and technical issues related to annotation and tools.

1.3.3.8 Summary

Although the concept of corpus linguistics has existed for more than thirty years, there is a mostly de facto tradition when dealing with linguistic data. Computerized corpora actually

[28] "La grande tolérance des praticiens du TAL envers diverses approches théoriques, pourvu qu'elles donnent lieu à des applications efficaces, s'explique et se justifie très bien si on adopte le point de vue de l'ingénierie. Elle n'a pas lieu d'être dès lors qu'on se place dans une perspective de recherche scientifique." (Cori, 2008, p. 109)

[29] "La tentative de concilier ces deux exigences quelque peu contradictoires a pu ainsi être à la source d'un brouillage des enjeux épistémologiques." (Cori, 2008, p. 109)

[30] "'Quantitatif' et 'qualitatif' n'opposent pas des méthodes (sinon dans un sens général, non scientifique mais plutôt seulement technique) mais des types de données." (Loiseau, 2008)

shaped corpus linguistics as an up to date discipline dealing with experimental traces, seeing itself as an instrumentalized text science, and constructing new observables on language.

Corpus linguists face a blur concerning the very notion of corpus linguistics as well as a blend of methodological and theoretical issues in the corpora they use. There is an oppositional aspect in the very definition of the discipline, based on a legitimization discourse which has proved to be productive. There are also diverging corpus-based and corpus-driven approaches.

All in all, corpus linguistics fails to convince a large part of the research community of its existence as an independent discipline on the one hand, and of its ability to bridge the methodological and epistemological gap between linguistics, experimental sciences, and applied linguistics on the other. Still, the impact of technological changes on the discipline keeps increasing in the course of time.

1.3.4 Enter web data: porosity of the notion of corpus, and opportunism (from the late 1990s onwards)

1.3.4.1 Technology, scientific policies, and technosciences as a background change

The notion of technosciences Reality as we know it today often manifests itself through technology, such as remote archiving, organizational sources in databases or automated processing chains. The experimental field itself comes equipped with a whole range of technologies, it remains editable and malleable long after its entry or registration and becomes a resource.

The background existence of what is increasingly called the technosciences plays a role in the mediation by technology, of which it can be said that, in the end, they offer much more than just a preservation or classification of information.

There are three main aspects of technosciences as Bensaude-Vincent (2009) describes them.

First, there is the emergence of scientific policies and funding agencies, that is to say for example the mutualization/sharing of political and financial means on a larger scale than before, as well as the distribution of funding on a competitive basis to fully-planned projects. This change is accompanied by the rise of a logic of profitability concerning results in the short-term, which plays a role from the selection of projects onwards.

Second, in connection with this development, research is increasingly oriented towards applicability, which gives technology the role of an indispensable tool in the production of knowledge and a place in the foreground. It is not necessarily about in-depth understanding of the phenomena anymore, but it is about making profitable use of nature or reality.

Third, knowledge and scientific fields are rearranged, most notably through the notion of convergence of theories and disciplines. Examples of this latter trend are present in several forms. The concept of “unification theory" in physics is an example that takes the form of a whole disciplinary field. The development of sciences of complexity as a fusion of different disciplines and theoretical models is another example. So are public policies (research funding and the underlying administration) the existence of the IXXI (complex sciences institute) in Lyon, devoted to the study of complexity in its various forms, as well as many “research clusters" in which the convergence of resources and scientists is crucial.

According to Gilbert Hottois, technosciences replace reality by a substrate. That which is (re)producible, manipulable, and/or transformable is actually considered to be reality, instead

of that which is visible, intelligible, or comprehensible.[31] The first is typical for the latest developments in sciences, while the second is the paradigm which used to be dominant and which is now supposed to be relegated to the background. In sum, according to Hottois, there is special emphasis to be put on interoperativity as a criterion of reality in the creation and utilization of resources.

Blind reason, Leibniz, and the age of cybernetics For a study of the philosophical origins and implications of the operative model, combinatorics, and "blind" science, see Barbaresi (2012). In this article, I shed light on the relationship between what Hottois calls "operative techno-logy" (in a functional sense) and the origins of this notion, dating back, according to him, to the calculability of signs by Leibniz, who writes about this particular combinatorial way of gaining knowledge that it is "blind" (*cognitio caeca vel symbolica*).

On the one hand, that which is visible plays a major role in philosophy, from the etymological meaning of "idea" to the examples used by philosophers and the association of light and reason. On the other hand, Leibniz had a great influence on the development of information systems and cybernetics, as Norbert Wiener for instance refers to him as a inspirational thinker.

Thus, using blind reason as a leading clue may be a productive way of thinking modern technology, be it with the foucauldian notions of dispositives and of the very realization of reason as a machine, with the criticism of Heidegger, who sees cybernetics as a systematic way to control the living, or with the techno-sciences, by Hottois or by Henry for instance.

From technosciences to interdisciplinary science However, if one thinks in Hottois' terms of the changes that have occurred since the 1990s in the way texts are accessed and analyzed, i.e. the existence of digitized, manipulable texts rather than simply texts seen as mere sources, this evolution makes sense beyond the framework of a mathematical view of language.

> "Science is increasingly interdisciplinary: the boundaries between previously distinct fields are blurring as ideas and tools are exported from one discipline to another. These shifts challenge the way that science is funded, conducted, communicated, evaluated and taught." (The Royal Society Science Policy Center, 2012, p. 37)

Combined with a growing empirical trend in linguistics (see above p. 21), the shift towards interdisciplinary science questions the established methodology of introspection or high-quality, limited corpora with new practices coming from computer science, especially concerning NLP and computational linguistics, and to a lesser extent linguistics. The "big data" and its afferent "more data is better data" paradigms are such examples.

1.3.4.2 On the notion of (big) data: background tendency of epistemological change?

Definition of data A comprehensive and trustworthy definition of data is given by a committee of selected researchers in their report for the Royal Society (The Royal Society Science Policy Center, 2012):

[31] "Est réel ce qui est (re)productible, manipulable, transformable et non plus le visible, l'intelligible ou le compréhensible." (Hottois, 1984, p. 62)

> "Data are numbers, characters or images that designate an attribute of a phenomenon. They become information when they are combined together in ways that have the potential to reveal patterns in the phenomenon. Information yields knowledge when it supports non-trivial, true claims about a phenomenon. For example, the numbers generated by a theodolite measuring the height of mountain peaks are data. Using a formula, the height of the peak can be deduced from the data, which is information. When combined with other information, for example about the mountain's rocks, this creates knowledge about the origin of the mountain. Some are sceptical about these distinctions, but this report regards them as a useful framework for understanding the role of data in science." (The Royal Society Science Policy Center, 2012, p. 14)

Thus, data can be seen as the bare scientific material, or unfiltered experimental results, while information stems from a conscious aggregation of data. This is a useful reminder to bear in mind: data alone cannot pretend to be scientific reasoning. What's more, data is not really useful if it does not come with metadata:

> "To be interpretable, data usually require some contextual information or metadata. This should include information about the data creator, how the data were acquired, the creation date and method, as well as technical details about how to use the dataset, how the data have been selected and treated, and how they have been analysed for scientific purposes. The preparation of metadata is particularly onerous for complex datasets or for those that have been subjected to mathematical modelling. But metadata are indispensible for reproducing results." (The Royal Society Science Policy Center, 2012, p. 14)

Thus, the report quoted here makes clear that data has to be both carefully prepared and carefully exploited, combined, or extracted in order to pave the way for a fruitful study.

The big data paradigm Concerning the size of web corpora in and of themselves, they can be considered to be part of the big data paradigm from the point of view of traditional linguists. In fact, what Baroni and Ueyama (2006) calls "large" is already (in 2006) a corpus in excess of 1 billion tokens (p. 33). Such a size is several orders of magnitude bigger than what was imaginable in the 1980s for example, when corpora in digitized form flourished.

Starting from this definition of large, one may expect corpora which are called "very large" to be one order of magnitude bigger, i.e. include more than 10 billions of tokens. From a sheer size point of view, such corpora are no distant goal, they are now perfectly accessible with decent hardware and processing architecture. However, there are striking changes in the nature and internals of text collections as they increase in size. Besides the theoretical and practical issues discussed throughout this document, there is a shift from carefully selected, meaningful linguistic material to "big data", an often ill-defined notion, whose interest in the context of web data and technosciences is described in this section.

According to (boyd & Crawford, 2012), the real nature of the big data paradigm is counter-intuitive, as it is not about size itself but about a whole dispositive surrounding the data:

> "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets." (boyd & Crawford, 2012, p. 663)

What links their analysis to the technosciences described above is the fact that they also consider Big Data as a "cultural, technological, and scholarly phenomenon".[32] According to them, it grounds on an interplay of several factors, from technological opportunities to nearly subconscious hopes and beliefs, which they reference as follows:

> "(1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
> (2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
> (3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."
> (boyd & Crawford, 2012, p. 663)

While this description is very general and applies to many different situations, it is perfectly acceptable concerning web data in linguistics. The latest technological developments (see p. 31), the search for patterns in language (see p. 41), and the belief that "big is beautiful" already mentioned by Habert et al. (1997) (see p. 35 for a discussion) are three elements that are to be found in the web corpus linguistics community. Nonetheless, I do not believe that the mythological part strictly applies to researchers who are usually deeply concerned by objectivity and accuracy (for instance). It may be present as a tendency in the background, more visible among the language technology community.

Is it really that much of a change? While there is undoubtedly a change of scale, one may also argue that it does not imply a change in methodology and practices per se. The big data paradigm cannot free itself from an existing tradition as well as from concrete technological achievements, as (Crawford, Gray, & Miltner, 2014) explains, taking the LHC (Large Hadron Collider) as example:

> "Big data is neither new nor free of the technical challenges raised since the emergence of supercomputers. Just as the Large Hadron Collider, the poster child for truly large data projects, has turned to magnetic tape as its best data storage solution, big data contains the techniques, artifacts, and challenges of older computational forms. Older concerns – technical, epistemological, and ethical – haunt the domains of big data." (Crawford et al., 2014, p. 1664)

Additionally, the belief that datasets are somehow going to deliver an adequate interpretation of themselves without any required intervention is fundamentally wrong, as the authors point out:

> "Interpretation is at the center of data analysis. Regardless of the size of a data, it is subject to limitation and bias. Without those biases and limitations being understood and outlined, misinterpretation is the result. Data analysis is most effective when researchers take account of the complex methodological processes that underlie the analysis of that data." (boyd & Crawford, 2012, p. 668)

[32] (boyd & Crawford, 2012, p. 663)

In short, research data should come with a manual (for documentation issues see p. 53). Even a clear trend towards self-interpreting data is to be taken with a pinch of salt, if not with a certain suspicion.

Big data applied to linguistics The potential primacy of quantity over quality is certainly the most striking shift accompanying the advent of large digital corpora. While it originally comes from other disciplines as well as from a favorable background of technosciences and big data described above (see p. 31), it has a real impact on the way linguistic research is conducted, not only research in computational linguistics.

The increasing importance of machine learning in both computational linguistics as well as in NLP (on the applicative side) gives a good grasp of this phenomenon. There are computational linguists who claim that both the technological evolution as well as the trend toward big data change the face of the discipline substantially:

> "As a result of improvements in computer technology and the increasing availability of data due to numerous data collection efforts, the data-intensive methods are no longer restricted to those working in affluent industrial laboratories." (Church & Mercer, 1993, p. 3)

Thus, contrary to Crawford et al. (2014)'s thesis, according to which there are old concerns between modern views, approaches, and technologies, there has been a branch of computational linguistics since the beginning of the 1990s feeding bigger data sets than ever before to data-intensive tools (mostly machine learning techniques, see for instance chapter 2) which either did not exist or were too costly to use, and which lean towards the applicative side.

The main effect of this approach is to downgrade the notion of corpus to the concept of (big) data as well as to favor work on the size of a set rather than on its quality, which is contrary to the tradition of corpora so far.

> "The wheel has turned, and now statistical approaches are pursued with great confidence and disdain for what went before. In a recent meeting, I heard a well-known researcher claim that the field had finally come to realize that quantity was more important than quality." (Kay, 2005, p. 436f)

All in all, in the massive aggregation of data that defines the big data paradigm, it seems that NLP, similarly to other disciplines, favors data quantity over data quality, which raises questions regarding the arguments of this approach.

1.3.4.3 Does more data mean better data?

The now well-known aphorism stating that more data is better data stems from the statistical machine translation community at the beginning of the 1990s, and more precisely in a joint article by Church and Mercer (1993):

> "It would appear that 'more data are better data', at least for the purpose of finding exemplars of words like *imaginable*." (Church & Mercer, 1993, p. 18)

Within the corpus design sketched so far the statement by Church and Mercer (1993) is definitely opportunistic. Regarding web scale data it raises many questions as to the relevance to linguistics and the practicability of corpus hypertrophy.

An increasingly popular paradigm Be it as a mere example collection or as a basis for quantitative analysis[33], bigger corpora seem to be more attractive. In fact, the popularity of bigger corpora seems to have been attested more than ten years ago[34].

"Size matters in NLP", claim Baroni and Ueyama (2006), and by NLP the authors also mean computational linguistics, as well as linguistics itself by extension:

> "Size matters also in other more theoretical fields of linguistics. For example, Mair has shown that the Web, unlike the BNC, is large enough to allow a full study of the grammaticalization of *get* as a passive in English." (Baroni & Ueyama, 2006, p. 32)

Here it is necessary to specify that "large", "huge", or even web-scale corpora do not necessarily mean open-ended corpora. The corpora can absolutely have a fixed size, which can be "as much text as possible", inverting the relation to statistical errors of Biber (1993): size is not a problem anymore but now the social and geographic diversity of the speakers is difficult to estimate.

That is were statistical approaches kick in, which back the popularity of large corpora, and to a greater extent big data, in linguistics.

Statistical approaches in linguistics: patterns and norms In a probabilistic perspective, the bigger the sample, the closer to the language population it is, as patterns emerge on a very large scale[35]. This data opportunism adopts another perspective on representativeness by claiming that larger corpora are potentially more representative.

This trend is present in corpus linguistics, where Perkuhn et al. (2012) consider the search for regular patterns as a major task of the discipline. This attention to not only the rules but also the norms can be used to justify corpus linguistics:

> "Traditional linguists have, for a long time, overlooked the fact that our utterances are much less the result of an infinite linguistic creativity, they are endless reiterations and variations of formulaic ready-mades. [..] Only corpus analysis makes us aware of such units of meaning. That meaning only emerges when single words co-occur with other words and are embedded in a wider context, inside a text embedded in a discourse, is something we have realised only with the advent of corpora and adequate software." (Teubert, 2010)[36]

Thus, according to the proponents of distributional approaches in corpus linguistics, the availability of new resources has drawn the linguists' attention to other units of meaning such

[33] "Neben der Verwendung von Korpora als eine Art 'Beispielbank' und rein qualitativen Analysen wird auch immer häufiger quantitativ gearbeitet." (Lüdeling, 2006, p. 28)

[34] "There was a live debate held in Oxford between prominent advocates of the two corpus design philosophies Quirk aided by Leech speaking up for the balanced corpus vs. Sinclair and Meijs arguing for the open-ended monitor corpus. Oral tradition has it that the debate was decided by the audience in favour of the Sinclair team."(Váradi, 2001, p. 591)

[35] "La conviction sous-jacente est que l'élargissement mécanique des données mémorisables (les centaines de millions de mots actuelles deviendront à terme des milliards) produit inévitablement un échantillon de plus en plus représentatif de la langue traitée. Si l'on n'arrive pas à cerner précisément les caractéristiques de l'ensemble des productions langagières, il ne reste qu'à englober le maximum d'énoncés possibles. À terme, la nécessité de choisir finirait par s'estomper."(Habert, 2000)

[36] The idea of "formulaic ready-mades" is related to the theories of Zellig Harris on language.

as collocations or co-occurrences, in short to patterns for which the study of lexicography or syntax for example is only partially relevant, because they are mostly approached "agnostically", for instance in terms of (co-)frequency.

In fact, the statistical approach to language is mostly different in nature from established grammatical and linguistic traditions:

> "A large corpus is most useful for studying less frequent items or, crucially, the macro-patterning of language that is not amenable to intuition and is ignored by grammatical tradition, and that can only be seen when many instances of relatively long sequences of items are brought together." (Hunston, 2008, p. 166)

The focus on "macro-patterning" is relatively new, it has probably been studied since the beginning of the 1990s. That said, the work of Quirk on the SEU described above (see p. 16) also focused on discovering regularities. A change of scale and new statistical tools make this paradigm shift possible, from mere intuitions to the challenge of the validity of intuitions compared to the estimated power of corpus data.

The precursors of such a paradigm, Church and Mercer (1993), highlight the need for such a change of scale, as well as the necessary corrections which are to be applied to the data:

> "Only a large corpus of natural language enables us to identify recurring patterns in the language and to observe collocational and lexical restrictions accurately. [...] However, in order to make use of this evidence we have to find ways to compensate for the obvious problems of working with unbalanced data." (Church & Mercer, 1993, p. 19)

In fact, Tanguy (2013) remarks that in the case of huge datasets, such as the Google N-gram dataset, many correctives are necessary.

Big data approaches to human language: less cleanliness, more data To be precise, an advantage computational linguists see in web data resides in the mere quantity of available occurrences, in the sense that more unclean data seems to be better than less clean data (Halevy, Norvig, & Pereira, 2009), which could per se enable statistical methods such as language models to obtain better results (Bergsma, Lin, & Goebel, 2009), regardless of data quality.

The interesting fact concerning this view is that in some cases it even tends to trump the tradition of linguistic annotation, which has not even indisputably established itself in linguistics yet.

> "The first lesson of Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available." (Halevy et al., 2009, p. 8)

In this respect, it is symptomatic that Halevy et al. (2009) acknowledge that massive, unfiltered, and linguistically unprocessed N-gram models come with a decrease concerning the quality of data, while highlighting at the same time the intrinsic value of large dataset:

> "In some ways this corpus [the Google N-grams] is a step backwards from the Brown Corpus: it's taken from unfiltered Web pages and thus contains incomplete sentences, spelling errors, grammatical errors, and all sorts of other errors. It's

> not annotated with carefully hand-corrected part-of-speech tags. But the fact that it's a million times larger than the Brown Corpus outweighs these drawbacks. A trillion-word corpus – along with other Web-derived corpora of millions, billions, or trillions of links, videos, images, tables, and user interactions – captures even very rare aspects of human behavior." (Halevy et al., 2009, p. 8)

This kind of reasoning is comparable to the mythological hopes regarding big data which I detailed above (see p. 32), hopes which may be defensible in other domains of application of the big data paradigm, but not necessarily on language data, and not from a linguist's perspective.

That said, such approaches ground on the established fact that language is not random, as well as on justified Zipfian considerations which lead to thinking that even unclean data cannot be completely wrong with respect to the language patterns that are uncovered within massive data sets.

> "For natural language applications, trust that human language has already evolved words for the important concepts. See how far you can go by tying together the words that are already there, rather than by inventing new concepts with clusters of words. Now go out and gather some data, and see what it can do." (Halevy et al., 2009, p. 12)

However, the principle of linguistics seen as a data science without theoretical foundations is an idea that also has been criticized by corpus linguists as a methodological failure as well as a potential danger with respect to data accumulation.[37]

Additionally, the big data paradigm can be seen as an example of abstraction and massive processing where focus is on machine learning and afferent benchmarks, so that even if fine-grained details may be considered, they are put in the background in favor of more remarkable trends. The more abstract the processes are, the less observable is the loss in terms of detail provoked by the linguists or the so-called data scientists (Bergounioux, 1992).[38] In a way, the statement by Bergounioux (1992) concerning a loss on the side of (roughly) introspective linguistics compared to field linguistics is also valid for data-intensive approaches: even if the loss can be formalized and quantified in the case of data-intensive linguistics and introspective linguistics, where it might not even be conscious, it is still a considerable loss from the point of view of data scrutiny.

Conclusion: critical return to big data To conclude, one may say that a certain criticism is necessary concerning the big data paradigm, which is often taken for granted without questioning. According to (Crawford et al., 2014), the glorification of big data resembles a "mythological artifact" whose function is to advertise that scale and scalability are everything and that existing methodologies are obsolete and must give way to data-intensive processes.

> "The very term big data science is itself a kind of mythological artifact: implying that the precepts and methods of scientific research change as the data sets in-

[37] "Korpuslinguistik läuft nicht auf eine Senkung der theoretischen, sondern auf eine Hebung der methodischen Ansprüche hinaus." (Lehmann, 2006, p.26)

[38] "Plus les procédures sont abstraites, modélisables par graphes et algorithmes, moins la déperdition provoquée par le philologue ou le linguiste est flagrante (ce pourrait être une des raisons du prestige de la syntaxe dans les études linguistiques)." (Bergounioux, 1992, p. 6-7)

crease in size. Some big data fundamentalists argue that at sufficient scale, data is enough." (Crawford et al., 2014, p. 1664)

It is useful to bear in mind that big data is not a neutral state of things, but a paradigm, a framework, a series of ubiquitous discourses:

> "We argue that big data is theory. It is an emerging Weltanschauung grounded across multiple domains in the public and private sectors, one that is need of deeper critical engagement." (Crawford et al., 2014, p. 1664)

Once the status of big data is redefined, it is possible to use data-intensive methods in suitable contexts of linguistics. By "suitable", I mean provided there is something to look for using this methodology. It fact, not everything relevant to linguistics is integrated or apparent in a text, so that Kay (2005) raises an interesting question regarding the "unseen" properties of language:

> "Statistical approaches to the processing of unannotated text bring up the thorny philosophical question of whether the necessary properties of language are, in fact, emergent properties of text. Could it be that at least some of the facts that one needs to know about a text are not anywhere in it?" (Kay, 2005, p. 430)

In corpus linguistics, relevant "emergent properties" may be missing simply because of a lack of annotation on a precise level. It is a matter of data quality and linguistic processing to make it possible to "learn" features from a corpus and draw conclusions. From an experimental point of view, the ecosystem in which a corpus is made accessible is of paramount importance, while from a theoretical point of view, the most fruitful methodology is probably to "cross the approaches" (Habert, 2000) and to create a setting for theoretical as well as experimental reasoning.

1.3.4.4 Interest of web data for computational, corpus, and theoretical linguistics

General interest of web data in (corpus) linguistics and humanities As of today, web data are mostly used without restriction in computational linguistics and NLP, while more questions are raised in linguistics. In a research context which has seen the rise of the big data paradigm and of corpus linguistics, the Web presents almost naturally as a potential source, because due to its vastness and its diversity most of the large corpora built today are taken from the Web, as Tanguy (2013) acknowledges.[39]

In the field of linguistics, web corpora have been an established scientific object for at least ten years (see p. 44 for the emergence of web corpora). This also holds true for the field of computational linguistics, with the seminal work of Resnik and Smith (2003), following experiments started at the end of the 1990s.

Content-focused uses only become significant when large-scale indexed and analysed corpora become available (Ostler, 2008).

[39] "Le Web est vu comme un réservoir indifférencié de textes à analyser, indépendamment des grandes questions sur leur nature. Cette caractéristique des approches ultra-massives des données en TAL se retrouve également dans le mouvement actuel en linguistique de corpus visant la constitution de corpus génériques de plus en plus volumineux ; il se trouve que les corpus actuels les plus volumineux (et de loin) sont issus du Web." (Tanguy, 2013, p. 7)

A paragon of web data in digital humanities is the newly discovered research field of culturomics, i.e. the quantitative analysis of texts to discover linguistic and cultural trends, for instance in the use of web news by (Leetaru, 2011). The new paradigm behind these studies is that understanding language (e.g. literature) is not accomplished by studying individual texts, but by aggregating and analyzing massive amounts of data (Jockers, 2013). However, initial studies in culturomics have been criticized for not referring to relevant work in linguistics and language technology (Borin et al., 2013).

The fact that massive amounts of data are elevated to the rank of a research object per se shows the close ties between big data as it is praised in other disciplines and the emergence of opportunistic, practical linguists.

“Big data" and “Web linguistics" The interest of web data is partly linked with the big data paradigm, as it echoes the opportunistic tendency of a substantial part of the corpus linguistics community:

> “Historically speaking, collecting data has been hard, time consuming, and resource intensive. Much of the enthusiasm surrounding Big Data stems from the perception that it offers easy access to massive amounts of data." (boyd & Crawford, 2012, p. 673)

Bergh and Zanchetta (2008, p. 309) use the expression “web linguistics" to refer to “empirical language research based on textual material collected from the Web". Additionally, the rareness of a number of linguistic phenomena seems to justify the need for large datasets, even for web scientists seeing linguistics “from the outside":

> “Natural language is a very sparse domain, in that most sentences uttered or written occur once only or very rarely, and the giant scale of the Web provides a fascinating corpus for NLP reasoning." (Berners-Lee et al., 2006, p. 49)

On the linguists’ side, Hundt, Nesselhauf, and Biewer (2007) ask “why take the risk of using web/non-traditional corpora?" Their answer can be divided into scientific, typological and technical reasons. To sum up, the most appealing characteristics of web data for corpus, theoretical and computational linguists seem to be first opportunism and data availability, second the need for linguistic proof material concerning finer and/or rarer phenomena as well as fine-grained statistical approaches, and last the desire to be at the cutting edge of language evolution.

In the following paragraphs, these points are described more in detail.

Opportunism and serendipity First of all, web texts are taken into account because they are there, already digitized, available and exploitable, a simple but effective reason mentioned by (Kilgarriff & Grefenstette, 2003):

> “Language scientists and technologists are increasingly turning to the Web as a source of language data, because it is so big, because it is the only available source for the type of language in which they are interested, or simply because it is free and instantly available." (Kilgarriff & Grefenstette, 2003, p. 333)

In combination with that comes a practical “opportunistic" argument which was already present in the beginning of the nineties (cf p. 19): why dismiss corpora and results because they are not fully conform to an optimal, hardly reachable research methodology?

> “It would be short-sighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as ‘unreliable’ or ‘irrelevant’ simply because the corpus used cannot be proved to be ‘balanced’." (Atkins et al., 1992, p. 14)

Additionally, corpus evidence soon becomes unavoidable, particularly for linguists unused to such profusion.

> “Perhaps the single most striking thing about corpus evidence – for lexicographers brought up on a diet of citation slips – is the inescapability of the information it presents." (Rundell & Stock, 1992, p. 22)

All in all, the fact that corpora, instruments, and finally proofs are at hand fosters a certain opportunism among the research community, all the more when coverage seems acceptable, as is explained more in detail p. 41.

However, taking a closer look at more traditional linguistics and corpus linguistics, the evolution is better explained by Renouf’s list of major drivers in corpus development: science, pragmatics and serendipity (Renouf, 2007, p.29). Concerning web corpora the primary driver is serendipity, since the web is a way to find useful and otherwise unavailable linguistic data, while the secondary motivation is pragmatic [40], which is roughly comparable to Kilgarriff and Grefenstette’s explanation. The “scientific" component, which (Renouf, 2007, p.29) defines as follows: “the desire to undertake an empirically-based methodological cycle", is either not primarily or not at all associated with what linguists expect from web corpora.

Confronted with an “expanding universe" ((Leech, 2006), see p. 16), linguists have to answer the question whether is it necessary or useful to set boundaries at some point. The answer is often negative, for reasons that deal with the porosity of the notion of corpus as described above and a certain opportunism regarding research data. The opportunistic perspective on size is in fact that it solves problems. The proponents of larger corpora may argue, for example, that “size will automatically sort out all questions of ‘balance’ in the structure of the data." (Váradi, 2001, p. 589) That leads to the pattern-based approach of linguistics.

Coverage, rarity of phenomena, and regularity of patterns Computational linguists such as Baroni and Ueyama (2006) evoke more detailed reasons to come to the same conclusion. For instance, according to Zipf’s law, while the most frequent words are easy to grasp, data sparseness becomes an growing issue if the words to be studied are rarer. This is even more the case when word combinations are taken into account, be it mere patterns or linguistic structures.[41]

[40] “Web texts, on the other hand, are freely available, vast in number and volume, constantly updated and full of the latest language use." (Renouf, 2007, p.42)

[41] “Because of the Zipfian properties of language, even a large corpus such as the BNC contains a sizable number of examples only for a relatively limited number of frequent words, with most words of English occurring only once or not occurring at all. The problem of ‘data sparseness’ is of course even bigger for word combinations and constructions" (Baroni & Ueyama, 2006, p. 31)

Another argument is of a statistical nature, and satisfies the constraints of the “pattern-searching" corpus linguists who wish to even out irregularities by analyzing a larger language sample. In fact, such a corpus is expected to provide “a more sophisticated and reliable statistical analysis" (Biemann et al., 2013, p. 35).

> “The Web is a dirty corpus, but expected usage is much more frequent than what might be considered noise." (Kilgarriff & Grefenstette, 2003, p.342).

Furthermore, as a consequence of the “more data is better data" paradigm described above, and because the notion has never been indisputably clear, the very definition of “corpus" as not being a simple collection of texts (Biber et al., 1998) changes, so that the proponents of this approach see corpora in a completely different light than the tradition in linguistics would have it (see above p. 11 for a description).

In fact, concerning web data, “opportunistic" and “statistical" views on language coincide, because they give access to a larger world of regularities and pattern discovery, with respect to the following research goals. First, finding occurrences of linguistic phenomena which cannot be found in traditional corpora, for example what Tanguy (2013) calls the “Pichon conjecture" concerning the suffix -este in French. Second, replicating the language population to find statistical evidence with a satisfying precision, in order to develop language models or to calculate token frequencies for instance.

For one of the primary uses of corpora, that being lexicographic purposes, (see p. 6), larger corpora are interesting in order to find occurrences of rare, unattested phenomena, or even structures that were considered to be impossible (Tanguy, 2013):

> “Only large corpora may allow us to witness unusual cases of word-formation or confirm whether some phenomena are rare or totally unattested." (Durand, 2009, p.8)

Even more so, Baroni and Ueyama (2006) claim that “we are reaching a size at which, at least for certain tasks, negative evidence (non-occurrence of a form in a corpus) can be taken as a linguistically meaningful fact." (Baroni & Ueyama, 2006, p. 36) Thus, in the particular context of coverage, precision is more important than recall, which paves the way for frequently updated, open-ended corpora.

Providing an up-to-date and diverse overview of language Other reasons are mentioned, such as the desire to study other genres/registers as well as technical language or sub-languages, and the constant evolution of language (Baroni & Ueyama, 2006, p. 31).

For example, Hundt et al. (2007) state that the existing corpora used in studies of English focus on inner-circle varieties of English, i.e. countries in which English is the first or the dominant language, with no data for other varieties. Yet other reasons are linked to the emergence of new linguistic developments, for instance new text types and genres closer to spoken language. The evaluation of the impact of Internet on language change is a research object in itself.

In fact, web corpora allow for “inclusion of up-to-date material so that recent trends and developments can be tracked", according to Biemann et al. (2013, p. 35), who specify that most of the texts included in the BNC were produced between 1985 and 1993. There is a necessary delay between text production and inclusion in the corpus with projects such as the BNC, delay which does not necessarily exist for freshly compiled web corpora.

Last, web corpora offer a more “democratic" perspective on language in the sense that phenomena and acceptability judgments are not invented or selected by a single linguist anymore, but reflect actual utterances of a wide range of speakers, actually “a broader range of authors and genres (such as fan fiction, semi-personal diaries, blogs, and forum discussions) than can be found in traditional corpora" (Biemann et al., 2013, p. 36).

To sum up, it can be said that web corpora are of interest because they allow for “extensive coverage, variability, freshness and open-endedness" (Bergh & Zanchetta, 2008, p. 310).

Other examples of cases where (very) large corpora are needed to perform linguistic studies Bergh and Zanchetta (2008, p. 310) consider the Web as a “unique and powerful alternative for different forms of empirical language research". They see a triple potential for Web linguistics: “as a resource for learning about authentic language structure and use", “as a provider of raw material for DIY disposable corpora and more ’stable’ collections", and “as a test bed for the training of various software applications" (Bergh & Zanchetta, 2008, p. 325).

Biemann et al. (2013, p. 29) list particular cases where such corpora are needed. They can be divided into three main fields of interest: studies on word and multi-word expression levels (lexicography, phraseology), morpho-syntactic analysis, and last but not least statistical analysis (distributional semantics, language modeling), as the statistical view of language is a current research trend (see p. 41).

Concerning the studies of word and multi-word expression levels, these phraseological units can be studied using large corpora, most notably the rare or new ones, as well as neologisms, with a plus for web corpora concerning non-traditional registers. Concerning morpho-syntax, Biemann et al. (2013) mention “rare (morpho-)syntactic phenomena of high theoretical importance". Finally, concerning statistical studies, statistical data for very infrequent words, word co-occurrences for infrequent words, and (structured) n-gram counts for language modeling are mentioned.

All in all, rarity is a key feature of phenomena which are meant to be observed using large corpora. One may add exhaustivity and statistical significance as potential motivators.

1.3.4.5 Holistic “web as corpus" approach and access to web data

A questionable approach: Googleology In order to find examples for rare phenomena, and because it is easy to use, linguists sometimes turn to Google, which is then considered to be an “online corpus" (as opposed to “offline", text collections downloaded and processed in a single shot), where the web itself can be considered as a corpus.

> “In Theoretical Linguistics, researchers sometimes try to obviate limitations of available resources through Googleology. Especially when data on low-frequency or non-standard phenomena is needed, search engine queries (mostly using Google’s service) are used to look for single occurrences of some grammatical construction, or – even worse – result counts returned for such queries are used for more or less formal statistical inference." (Schäfer & Bildhauer, 2013, p. 4)"

For a number of reasons, the results of this Google-centered approach are scientifically questionable. The caveats are listed by Kilgarriff (2007), and later by Schäfer and Bildhauer (2013) for example. They can be split into the following categories: search engines are unadapted, their functioning is obscure, and the results are not reproducible.

First of all, search engines are not adapted for linguistic research, because they favor precision over recall, and what's more, they do so according to criteria which have nothing to do with linguistics. Search engines are also inapt to answer linguistically-founded queries because they offer no linguistic annotation of the documents. Additionally, features such as controlled wildcarding are not available, and the engines may expand and reduce search terms without being asked to do so and without giving any hint on the operation, which leads to the second category of problems.

In fact, the way search engines process queries and classify results is obscure and mostly undocumented due to the need to keep a competitive edge. The extent of spelling correction and language-specific settings for queries is unknown, as well as the ranking criteria. The latter do ground on the well-known PageRank algorithm, but have been constantly adapted and modified since the beginning, in ways that are once again undocumented. What's more, drastic changes in ranking may happen without the knowledge of the users. All these factors let search engines become what Bergh and Zanchetta (2008) call "a baffling experience" (p. 316).

Last, the querying process is not reproducible, because search engines adapt the ranking of the results according to user location and settings and also because they return estimated total counts, which may vary quite frequently. Finally, even if what is indexed by the search engines could be labeled a corpus, the very nature of the web is such that the sample which the engines take into account is bound to change at a fast pace.

Nonetheless, cases have been reported where specific functions of search engines could be used with satisfying results, for example in regional dialectology (Grieve, Asnaghi, & Ruette, 2013). There are precise reasons making such a study possible, most notably the size of the US territory and the existence of numerous local newspapers, which allow for a relatively low granularity of results. Detailed statistical analyses are performed in order to make sure that data is clean enough and that results are significant. Clear cases of regional variation can be distinguished, but the method does not allow researchers to draw conclusions about negative or unclear results.

All in all, for the reasons detailed above (the list is not exhaustive), there is a broad consensus among web corpus specialists that Googleology as it has been called by linguists cannot be considered a valid scientific approach. Sadly, this point of view has not reached the entire linguistics community yet, and it is hard to estimate to what extent it has, since search engines may be used for exploration or verification purposes only, and thus not publicly referenced as a source. Still, because of the demonstration above, Googleology will not be mentioned further in this document.

Emergence of the "web as corpus" paradigm Due to technological progress and to the popularity of data-intensive approaches, there has been a tendency in the 1990s and later towards dynamic and open-ended corpora (Renouf, 2007), with the emergence of the "web as corpus" paradigm in 1998, a paragon of open-ended corpora.

Web data collection also became possible at a time where a general development is taking place in linguistics, towards the acceptance of corpora as valuable source:

> "It is accepted that the compilation of natural language corpora is a valuable use of research resources"(Sampson, 2000, p. 1350)

The end of the nineties witnessed the advent of corpus-based approaches as largely accepted scientific methodology, whereas the early 2000s saw the beginning of the "web as corpus"

paradigm:

> "We seem to be witnessing as well a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use Internet data" (Joseph, 2004)

The "web as corpus" paradigm is a typical "opportunistic" view of corpora and corpus construction. It redefines the notion of corpus in order to adapt it to new possibilities. Thus, the question whether a corpus is a mere collection of texts or whether there has to be scientific intention and methodology behind it is answered in a different manner:

> "A corpus is a collection of texts when considered as an object of language or literary study. The answer to the question 'is the Web a corpus?' is yes." (Kilgarriff & Grefenstette, 2003, p. 334)

Access to web data At the beginning of the 2000's, in perfect compliance with the "Web as corpus" paradigm, the first tools developed by linguists were designed to query search engines and refine their results in a more suitable fashion, for example "quick word in context" (KWIC) concordancers such as KWiCFinder or WebCorp (Renouf, 2007).

> "Webcorp Live (Kehoe & Renouf, 2002) is a Web-based tool that allows to use the web as a concordancer as the results taken from Google are presented in a more linguist-friendly manner. However, Webcorp is not a crawler as the whole web pages cannot be downloaded for further use. This also excludes statistical analysis of the results or further linguistic transformations like tagging or lemmatization. The Webcorp Linguist's Search Engine gives access to a set of preconfigured and preanalyzed English language webdata." (Gerdes, 2014, p. 269)

Although it was a practical way to harness the power of a search engine, this kind of approach soon fell short, first because of the lack of linguistically relevant information:

> "Linguist-friendly interfaces to search engines, such as WebCorp, while useful in that they reformat/reorganize the data returned by the search engine in ways that are more conductive to linguistic research – KWIC display, frequency lists – are not providing any more information than what is provided by the search engine, and, thus, they present the same problems" (Baroni & Ueyama, 2006, p. 32)

Secondly, the tools refining search engines' results also fell short regarding consistency, reproducibility and ultimately validity of the results, as it became clear that this kind of information was far from being unbiased and as the debate on Googleology emerged (see above p. 43).

While a few computer linguists discovered early on the potential of the web for corpus and computational linguistics purposes, web texts really began to be used by a growing user base with the Special Interest Group of the Association for Computational Linguistics on Web as Corpus[42] and web as corpus frameworks such as WacKy project (Baroni, Bernardini, Ferraresi,

[42] ACL SIGWAC
https://www.sigwac.org.uk/

& Zanchetta, 2009) and its BootCaT method (Baroni & Bernardini, 2004). This approach is not limited to English, it has been used for other major world languages (Baroni et al., 2009; Kilgarriff, Reddy, Pomikálek, & Avinesh, 2010).

The “sketch engine" (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) and the “linguist’s search engine" (Resnik & Elkiss, 2005) are both designed to be accessible tools for linguists. This kind of software offers a straightforward approach to web text collection, and the ease of its use is probably a determining factor in the adoption by most researchers.

In linguist-friendly interfaces, the Web is both media and medium, both data and channel to access the data, which means that it is easier to work with an existing corpus provided an efficient interface is available:

> “The Web is not only a source of free, already-digitized data (doing away with the need to scan or key in texts). It is also the channel through which we access the data. For today’s lexicographer, the usual working method is to view and analyse corpus data online, so we no longer need to install either the search software or the corpus itself on our own machines." (Rundell, 2008, p. 26)

Besides, concerning derivatives of web corpora, the Google Web Ngram project (Brants & Franz, 2006) is a milestone on the way towards an image of language as found on the web, since it is widely used across disciplines, even more so since the release of the “Web 1T 5-gram database", which is considered to be “the largest publicly available resource derived from a web corpus" (Biemann et al., 2013, p. 28), with about 1 trillion words of English text. Even if restrictions apply concerning the content, it has been used for linguistic purposes:

> “However, it is not distributed in full-text form, but only as a database of frequency counts for n-grams of up to five words. Throughout, n-grams with fewer than 40 occurrences were omitted from the database. Due to these restrictions, this resource is of limited use for linguistic purposes, but it can be – and has been – applied to certain types of analyses such as collocation identification."(Biemann et al., 2013, p. 28)

Recently, there has been a growing interest in the research community in web texts taken from the CommonCrawl project[43], which consists of retrieved web pages made available at little or no cost. Researchers in the machine translation field have started another attempt to outsource competence and computing power to find parallel corpora (Smith et al., 2013). For a critical review of this approach, see chapter 4.

To conclude, according to Mair (2012), there are currently two major trends to access web data, “direct" and “offline" use:

- Direct and systematic use (e.g. discourse analysis, culturomics). However, the database is unstable and search engines do not allow for sophisticated queries.
- Compiling offline corpora from Web sources, in the form of a representative "mini web", these “technological" issues are tackled by ACL SIGWAC.

In linguistics and computational linguistics, both approaches coexist, although, as shown concerning Googleology (see p. 43), specially compiled corpora are preferable.

[43] http://commoncrawl.org/

1.3.4.6 Summary

As I have described concerning the empirical trend of linguistics (p. 21), the emergence of interdisciplinary work, and technosciences (p. 31), the statistical view on language is becoming more and more popular, for instance with the study of language patterns. In that context, the availability of web data, i.e. more language data than ever before, fills a gap (Tanguy, 2013). Consequently, large corpora are en vogue, and many researchers hope that the sheer mass of data may balance a lack of data quality.

Tanguy (2013) sees three main phases in the use of web data in linguistics.[44] First, “wild" usage of search engines through ad hoc software. Second, a balance between the enthusiasm of researchers and the cost of massive querying. Third, the current phase of retreat from search engines, with the wish to make static or derived resources available.

However, from a general point of view the segmentation into three different phases is not so clear. Access to web data occurs through concurrent approaches. In fact, even nowadays and mostly because of the lack of alternatives, search engines are used as a convenient black box by language learners for instance, even if a proper interpretation of the results is not straightforward (Boulton, 2013). Concurrently, there are researchers building their own offline corpora and gathering expertise in web crawling and corpus processing (cf chapter 3). Thus, there is a continuum between “wild", heedless approaches based on black boxes, researchers working on derivatives such as n-gram models, and expert approaches where ad hoc corpora are built, replicating the global functioning of search engines or for specific purposes.

The very notion of corpus is adapted to embrace a world of opportunism and serendipity, while in parallel corpus work becomes more and more accepted in linguistics. Not only the availability, but also the coverage in terms of registers or languages convinces researchers outside of the corpus linguistics community. As Rundell (2008) sums up:

> “The arrival of the Internet, and its extraordinary growth, has put at our disposal more or less infinite quantities of digitized text in a wide range of registers, and this has become the raw material for contemporary corpora (in English and many other languages)." (Rundell, 2008, p. 24)

In this section, a history of corpus construction has been sketched and basic tendencies of corpus linguistics have been identified. Now that the scenery is set, the next section will deal with the changes in the notion of corpus and corpus contents from a methodological point of view.

[44] “On peut distinguer trois grandes phases : l’utilisation « sauvage » de ces moteurs par des programmes spécifiques, suivie d’une période de compromis entre l’enthousiasme des chercheurs et le coût (pour les moteurs) des interrogations massives, et enfin une période actuelle de retranchement des moteurs de recherche, qui privilégient dans le meilleur des cas la mise à disposition de la communauté scientifique de ressources statiques ou de produits dérivés, comme des bases de données de séquences de mots (n-grammes)." (Tanguy, 2013, p. 4)

1.4 Web native corpora – Continuities and changes

1.4.1 Corpus linguistics methodology

1.4.1.1 Web corpora as a magnifying glass

Not all of the methodology has to be changed simply because the texts are natively digitized and available on the Web. On the contrary, it is generally so that existing methods are transferred or adapted to the digital age:

> "Moreover, although scientists do routinely exploit the massive data volumes and computing capacity of the digital age, the approach is often redolent of the paper age rather than the digital age." (The Royal Society Science Policy Center, 2012, p. 16)

Indeed, to some extent the corpus linguistics methodology stayed the same for the first generations of corpora from the web. Methods used so far were reused, but for different research objectives:

> "It is likely that this first generation of linguist's search engines and the underlying web corpora will look like oversized versions of the corpora we know (billions of words rather than hundreds of millions of words), solving some of the sparseness problems of current corpora, but still far away from exploiting all the dynamic linguistic potential of the web." (Lüdeling, Evert, & Baroni, 2007, p. 21)

In that sense, one may speak of a "magnifying glass", as described by Hundt et al. (2007), exemplifying the constraints of "traditional" corpus construction as well as the weaknesses of corpora before and after the web.

> "The ongoing discussion and the articles in this volume show that many corpus linguists are still very much concerned with issues such as representativeness, structure, balance, documentation and replicability, especially when it comes to the web as a source of information. These issues now have to be re-addressed from a new angle – it could be argued that the challenge of using the www in corpus linguistics just serves as a magnifying glass for the methodological issues that corpus linguists have discussed all along." (Hundt et al., 2007, p. 4)

Because so much has been taken for granted in web corpora on the one hand and so little accepted as a scientifically valid approach on the other, there is still much to look for, to annotate, and to prove regarding the linguistic use of web data:

> The Web is a "deceptively convenient source of data – with a 'corpus architecture' that is taken for granted and retrieval software that is considered a global default standard and therefore not worth thinking about" (Mair, 2012).

There is also still much to question regarding web corpus retrieval and architecture. In this section, I will discuss the changes and similarities between "traditional" corpora and corpora from the web. More specific issues are mentioned in chapter 3, most notably concerning content sources and web text selection, while solutions to the described problems are discussed in chapter 4.

1.4.1.2 Two types of corpora? Web as corpus and web for corpus

Definitions: Web as corpus and web for corpus According to Hundt et al. (2007), there are two different uses of the web in corpus-linguistic research: first what the authors call the “web as corpus" approach, featuring commercial crawlers and internet-based search engines, with either a heuristic methodology they call “data sniffling" or a systematic application that they call “data testing". Second, the web can be used as a source for the compilation of large corpora called “offline" because they may be accessible on demand through corpus management tools. The authors label this approach “web for corpus building".

As mentioned by Tanguy (2013), the first clear distinction between Web as and Web for corpus probably dates back to De Schryver (2002), who, in the case of African languages, suggested to distinguish between approaches which use the Web to build a corpus and others that consider the whole Web as one single corpus.

That said, even if the conceptual difference is pertinent, it would now be presumptuous to label one’s approach “web as corpus", because even the most frequently used search engines don’t cover the whole web. It is too large, hence they have to take shortcuts and exclude parts of it using efficiency rules. By way of consequence, only the “web for corpus" or "corpora from the web" approach has prevailed.

Additionally, the distinction between “online" and “offline" corpora is not currently justifiable either, if it ever was. In fact, even early web corpora using for instance the BootCaT approach (Baroni & Bernardini, 2004) were not meant to remain up-to-date. Large text collections require a considerable amount of time to process, and are usually crawled, processed and made available once and for all, for instance once per year like the COW project (Schäfer & Bildhauer, 2012). The large web corpus projects usually make data available through a querying interface for demonstration or research purposes, so that they can be described as offline corpora accessible online.

A real difference exists between so-called “general" and “specific" or “specialized" corpora, a well-known characteristic often used in corpus typology (for a general typology see above p. 13).

> “The most basic distinction is that between general corpora and specific corpora. The former intend to be representative and balanced for a language as a whole – within the above-mentioned limits, that is – while the latter are by design restricted to a particular variety, register, genre,..." (Gries, 2009, p. 1232)

As the difference in corpus composition reflects divergent research goals described in the following paragraph, I prefer to speak of “general-purpose" corpora rather than “general" corpora, in order to focus on their inclusive character as well as on the “opportunistic" nature of most web corpora. I also tend to avoid the qualification “general language" for web corpora, since the nature of their content is subject to great variation, and it is unclear whether it overlaps with the concept of “general language" as termed by most linguists.

Research goals On the one hand there are all-purpose, “one size fits all corpora", allegedly fast and easy to construct and useful for “big data" approaches. On the other there are specific corpora with controlled text inclusions and possibly rich metadata, built with particular research goals in mind, such as variation-aware approaches which take production conditions into account (like in classical reference corpora).

Atkins et al. (1992) already saw two major classes of uses for corpora, new tools searching large datasets being opposed to well-known corpus content which is conventient for performing tests on:

> "The corpus as a large-scale, but heavily-diluted source of data, which new techniques are enabling us to sift; and the corpus as a testbed, composed of representative if mostly undifferentiated material, good for testing or training an automatic device under development." (Atkins et al., 1992, p. 28)

Obviously, large general-purpose web corpora grant access to a world of the unknown where robustness and efficiency are key characteristics, whereas specific ones enable to find linguistic evidence at a lower cost, following established techniques while remaining comparable to previous results.

Control Another divergent element deals with controlling experimental parameters, and concerns the representativity of language as a whole and/or particular varieties. On the one hand, there are cases where representativeness can be assessed or controlled and where the corpus is known to be adequate for observing given phenomena. On the other hand, there are cases where size and content diversity account for a better coverage and allow for better use of statistical and heuristic processes[45].

It is also possible to identify two different types of access to corpora (Hundt et al., 2007), with on the one hand search-engine based access and on the other proper corpus building. In the first case the user has very little knowledge about the content, corpus building and query results are impossible to reproduce, and there is an inbuilt local bias of crawlers. In the latter case, a mastered corpus building allows for more control (including what the designers want), accessibility (standard software tools), and level of analysis (linguistic annotation procedures).

In-between, evolutionary corpus design The contrast between general-purpose and specific corpora may evolve, as these categories are not impermeable. It is possible to find corpora that are in-between, or transferred from one to another thanks to later developments in corpus design.

First of all, an opportunistic corpus is generally considered to be a specialized corpus, which is why "opportunistic" is not a special category in corpus typology (Lemnitzer & Zinsmeister, 2010, p. 107). However, depending on corpus composition, but also because of the existing different definitions of corpus, a very large corpus from the web may be considered either an opportunistic and thus specialized corpus or a general-purpose one. In fact, it could be labeled as in-between.

Second, specific corpora can also be sampled or constructed as subsets of large corpora for specific purposes (Eichinger, 2006, p. 5), provided the larger corpora come with enough metadata to satisfy the constraints of this construction process. Adaptable corpora that can suit the needs of different research projects are also called "dynamic" corpora (Hundt, 2008). It is a useful characteristic to be able to adapt the definition of the word corpus and the subsequent corpus design decisions and to apply them to reshape a given corpus:

[45] "Zum einen kann man sich ein Korpus vor- und zusammenstellen, das im Hinblick auf die vermutete Repräsentativität für bestimmte Phänomene oder für bestimmte Varietäten zusammengestellt ist. [...] Zum anderen kann man sich vornehmen, über die schiere Größe und Vielfalt von Korpora die Reliabilität von Korpora zu erhöhen."(Eichinger, 2006, p.4)

> "With an increasing array of general language corpora that will obviously not be perfect fits for cross-corpus comparisons, a flexible corpus structure that allows users to adapt the size and composition of the corpora will be a convenient solution." (Hundt, 2008, p. 174)

Following this assumption, the builders of the German IDS corpus described it as being an "original sample" (*Urstichprobe*) (Perkuhn et al., 2012), which is genuinely "general-purpose" in the way that it is meant to generate subcorpora corresponding to specific views on for instance representativeness/balance or text genre.

The notions of "main" and "reserve corpus" seem to have first been implemented by the COBUILD project in lexical computing (Johansson, 2008, p. 42). They are now also used by other reference corpus projects such as the DWDS (Geyken, 2007):

> "In addition to the Kerncorpus, the DWDS project also compiled a much larger corpus from electronic versions of daily and weekly newspapers of the 1990s. [...] This opportunistic corpus, the DWDS Ergänzungscorpus (supplementary corpus), comprises approximately 900 million tokens gathered in two million articles." (Geyken, 2007, p. 27)

The goal here is to address at least two types of users, among them "traditional" linguists who wish to perform studies on a balanced reference corpus, and "opportunistic" linguists who wish to include every valid and available resource.

Newspaper corpora may build so-called "monitor corpora" (Renouf, 1993) without fixed content. These corpora are supposed to follow language evolution.

Typology Following the typology described in section 1.2.3.4 (Lemnitzer & Zinsmeister, 2010, p. 103), table **??** summarizes typological differences between general-purpose and specific web corpora.

Criterion	Specific	General-purpose
Functionality	*a priori*	depends
Language selection	monolingual or parallel	usually monolingual
Medium	written texts	written texts
Size	limited	(very) large
Annotation level	possibly fine-grained	POS-tagging
Availability	depends	usually queries

Table 1.1: Synoptic typological comparison of specific and general-purpose corpora

1.4.1.3 Examples of corpora

General-purpose corpora Current general-purpose web corpora are what could be called "one size fits all" corpora, since they are built with a particular balance of size, coverage, and attention to detail. This balance is supposed to fit the needs of as many users as possible. The case of corpora developed in a company co-founded by Adam Kilgarriff is eloquent, since the business model of the company is based on an extensive user base, whose needs are to be addressed. For instance, new languages, more ways to classify corpus data, or a more detailed

interface are added. Only part of these internal processes are known, principally the gathering of corpora which is currently performed by the Spiderling software (Suchomel & Pomikálek, 2012).

Current tools such as the Spiderling (Suchomel & Pomikálek, 2012) or the COW (Schäfer & Bildhauer, 2012) projects start from seed URLs extracted from search engine queries, then use full-fledged language-focused crawlers and finally language-focused processing and optional refinement processes (cf next chapter for more details).

Specific corpora Seminal work on the topic of specific/specialized corpora is to be found in Hoffmann (2006). Hoffmann used a specialized corpus gathered on the Web to answer a particular research question: can Internet data be used as a basis for quantitative analyses of present-day English?

He built what he called a "specialized corpus of spoken data" made of publicly available CNN transcripts retrieved from the website `cnn.com`. He found that even if restrictions apply, it is completely valid to build a specialized corpus on these terms.

Whereas little has changed in the approach of specific corpora, design decisions and usages may be really different when it comes to general-purpose web corpora (see section below).

The Czech internet corpus: Web for "old school" balanced corpus The Czech internet corpus (Spoustová & Spousta, 2012) is a good example of focused web corpora built in order to gather an "old school" balanced corpus encompassing different genres and several text types.

The crawled websites were not selected automatically nor at random but according to the linguists' expert knowledge: the authors mention their "knowledge of the Czech Internet" and their experience on "web site popularity". The whole process as well as the target websites are described as follows:

> "We have chosen to begin with manually selecting, crawling and cleaning particular web sites with large and good-enough-quality textual content (e.g. news servers, blog sites, young mothers discussion fora etc.)." (Spoustová & Spousta, 2012, p. 311)

Finally, they divided the corpus into three parts: articles, discussions and blogs. What they did with mixed-content is not clear:

> "Encouraged by the size, and also by the quality of the texts acquired from the web, we decided to compile the whole corpus only from particular, carefully selected sites, to proceed the cleaning part in the same, sophisticated manner, and to divide the corpus into three parts – articles (from news, magazines etc.), discussions (mainly standalone discussion fora, but also some comments to the articles in acceptable quality) and blogs (also diaries, stories, poetry, user film reviews)." (Spoustová & Spousta, 2012, p. 312)

There are indeed articles and blog posts which due to long comment threads are likelier to fall into the discussion category. On so-called "pure players" or "netzines" the distinction between an article and a blog post is not clear either, because of the content but also for technical reasons related to the publishing software, such as content management system like

WordPress, which is very popular among bloggers but also sometimes used to propel static websites.

It is interesting to see that “classical" approaches to web texts seem to be valid among the corpus linguistics community, in a shift that could be associated with the “web for corpus” or “corpora from the web” approach.

The workflow replicates steps that are useful for scanned texts, with boilerplate removal somehow replacing OCR corrections. One clear advantage is the availability and quantity of the texts, another is the speed of processing, both are mentioned by the authors who are convinced that their approach can lead to further text collections. A downside is the lack of information about the decisions made during the process, which ought to be encoded as metadata and exported with the corpus, so that the boilerplate removal or the text classification process for example can be evaluated or redesigned using other tools.

1.4.1.4 Web corpora, language documentation, and less-resourced languages

Language documentation Language documentation can be defined “a lasting, multipurpose record of a language" (Himmelmann, 2006, p. 1).

> “There has been a tremendous upsurge of interest in documentary linguistics, the field concerned with the the ‘creation, annotation, preservation, and dissemination of transparent records of a language’ (Woodbury, 2010)." (S. Abney & Bird, 2010, p. 88)

Corpus construction can be considered to be closely related to language documentation, or even a subpart of it in certain cases such as for instance for the less-resourced languages. In fact, primary data first has to be found, then it has to be compiled and finally made available. The main difference between language documentation and most corpora is that the latter cannot pretend to document something which cannot be replaced by the acquisition of new data (Lemnitzer & Zinsmeister, 2010, p. 108).

Special attention to recording, processing, and preserving primary data Texts are more easily collected using the web now than before its existence, and the whole collection process is a lot faster than in other fields of inquiry using language documentation. As such, web corpora are much more inclined to follow mottoes such as “fail fast/early and fail often"[46]. Even if the research costs are substantial, it is conceivable to apply a trial and error methodology to web crawls or web corpus building in general.

However, it is desirable not to let practical concerns and results alone drive the research methodology. Web corpora need theoretical grounding in order to avoid creating “data graveyards", as much as language documentation processes do according to Himmelmann (2006).[47] In fact, language documentation is not a theory-free discipline according to Himmelmann (2006), nor should web corpora become such a purely experimental object:

> “Language documentation is not a theory-free or anti-theoretical enterprise. Its theoretical concerns pertain to the methods used in recording, processing, and

[46] Silicon Valley motto of unknown origin.

[47] “Without theoretical grounding language documentation is in the danger of producing ‘data graveyards’, i.e. large heaps of data with little or no use to anyone." (Himmelmann, 2006, p. 4)

> preserving linguistic primary data, as well as to the question how it can be ensured that primary data collections are indeed of use for a broad range of theoretical and applied purposes." (Himmelmann, 2006, p. 4)

Thus the theoretical concerns should focus on three steps: the recording, processing and preserving of linguistic data. If applied to web corpora, these steps could be first web crawling and web document retrieval, second preprocessing and inclusion in the corpus, and last linguistic annotation, encoding format and querying process.

Making web data available and accountable The focus on primary data is an important similarity between language documentation and web corpus construction. In both cases, it is crucial to deliver as many documents as possible, and the relative importance of documents is secondary.

> "The main goal of a language documentation is to make primary data available for a broad group of users. Unlike in the philological tradition, there is no restriction to culturally or historically 'important' documents, however such importance may be defined." (Himmelmann, 2006, p. 15)

The explicit concern for accountability expressed by Himmelmann (2006) is a key concept: transparent collection processes are preferable to invisible or black box processes. Moreover, proper language documentation relies on metadata and quality of data.

According to (Austin, 2010), documentation requires a scientific approach to several collection steps: first information capture, second data structuring, processing and analysis, third data archiving and preservation and last mobilization. The two first steps are highly relevant in this context, whereas the third and fourth are at least partially so.

The case of less-resourced languages The notions of "lesser-known", "low-resource", "minority", "noncentral", and "under-resourced" languages are found in the literature. These denominations put an emphasis on resources and language technology, and as such they do not necessarily overlap with languages considered large by way of their speaker count[48]. This conceptual flaw accounts for the diversity of situations encountered and the difficulty to find "one size fits all" solutions.

The interest for less-resourced languages stems from the interest for language documentation on the one hand, which has been increasing recently, as described above, and from the language technology and linguistic resources community on the other hand, which is particularly active:

> "In the last few years, there has been an increased interest among the language technology research community in developing methodologies that would minimize both the data requirements and the human linguistic expertise needed for the creation of linguistic resources and language technology tools." (Borin, 2009a, p. 4)

[48] "Large standard languages – those with numbers of native speakers in the hundreds and tens of millions and having a long tradition of writing – are not necessarily high- or even medium-density languages." (Borin, 2009a, p. 3)

Interest of web data for less-resourced languages There is a broad consensus among researchers concerning the idea that corpora from the web are a relevant way to build new resources, in a context where "the first half century of research in computational linguistics – from circa 1960 up to the present – has touched on less than 1% of the world's languages" (S. Abney & Bird, 2010). The need for tools is manifold:

> "We need tools for annotation, format conversion, spidering and language identification, search, archiving, and presentation." (S. Abney & Bird, 2010, p. 94)

Spidering –language identification and format conversion, to put it into the right processing order– are issues which will be addressed in the remainder of this document, while the other steps will only be marginally mentioned.

In the context of lesser-known languages and language documentation, the Web is acknowledged as a potential source, mostly due to the availability of material:

> "The web can be used as a source of (a kind of) corpora, for assembling them on the fly" (Borin, 2009b, p. 14)

In fact, the presence of texts on the Web enables researchers to gather them even if they know little about the language, the key being a proper trustworthy language identification system. The rest of the processing chain can most of the time be used as is, thus allowing for serial production of web corpora:

> "[The Web] allows fast and cheap construction of corpora in many languages for which no standard reference corpus such as the BNC is available to researchers. This set does not include only so-called 'minority languages', but also well studied languages such as Italian and Japanese." (Baroni & Ueyama, 2006, p. 32)

The Web has become far too large for linguists to pretend to cover a significant portion of web space. On the contrary, finding texts for languages that are not widespread is a real challenge. Concerning minority languages, there are two major restrictions to the process:

> "Of course, the languages must be written languages, and there must be a sufficient number of web publications in them. Of the two, the first is the more restrictive requirement, since only a modest fraction of the world's languages are written. Even written languages are quite unevenly represented on the web, however." (Borin, 2009b, p. 14)

Not all languages that fall into the scope of language documentation are written. Concerning the second restriction, i.e. the existence of web publications, the trend is quite positive, as internet access is increasingly popular and affordable around the world. Thus, despite the relative disparity between languages on the Web one may expect to find more and more language samples.

1.4.2 From content oversight to suitable processing: Known problems

1.4.2.1 Challenging the text criterion

The very characteristics of web documents may be a problem, for instance concerning efficient tools for automatic removal of web-specific formatting (Hundt et al., 2007, p.4), as well as the difficulty of defining the notion of text and what (Renouf, 2007) calls the “handling of web pages with their hotchpotch of more and less text-like texts".

Compared to well-known and in a way “controlled" corpora, the “heterogeneity and arbitrariness of the text" (Renouf, 2007) contained in web corpora may be questioned, as it is opposed to the ideal notion of text described by Atkins et al. (1992) (see p. 11).

First, not all web texts are of discursive nature.

Second, a considerable amount is much shorter than books or newspaper articles.

Third, web texts are not necessarily integral, due to the very nature of web pages (and the interlinking and content injection sometimes called the Web 2.0), or due to the fragmentary nature of popular text genres such as comments, microtexts, or follow-ups. As a consequence, web texts are often not “conscious products of a unified authorial effort", among other reasons because the possible authors are not aware that they are cooperating, and/or because there are no common writing guidelines, for instance on the homepages of blog communities, social networks, or online versions of newspapers, which aggregate content from different sources.

Last, despite the existence of well-documented guidelines and the enforcement of common editing rules, the texts on Wikipedia – one of the most frequently viewed websites in the world and home to a considerable amount of usable text – cannot be considered to be “stylistically homogeneous".

Another typical phenomenon for the web as it has evolved is that text often comes second to images, audio files, videos or multimedia content, both in terms of web page design and actual time spent on audio/video platforms or social networks, which are among the most popular and among the largest existing websites.

1.4.2.2 To what extent do new text types and genres lead to a lack of metadata?

Description of the problem Among the loosely-defined group of internet users there are new or so far unobserved linguistic populations. There are also spontaneous and less spontaneous uses, different intentions, a vast array of possible audiences and communicational goals which in sum account for corpora of a different nature than the canonical text corpora. From microtext by the elderly to sponsored fashion blogs, including machine-translated and automatically generated content, this diversity questions the good practices from pre-web times. Biber and Kurjian (2007) describes the problems in terms of language registers:

> “With most standard corpora, register categories are readily identifiable and can therefore be used in linguistic studies. However, research based on the web lacks this essential background information. [...] The fundamental problem is that we have no reliable methods for identifying the kinds of texts included in a general web search. In fact, there is an even more basic underlying problem: we do not at present know what range of registers exists on the web." (Biber & Kurjian, 2007, p. 111–112)

In other terms, it is what Bergh and Zanchetta (2008) called the “heterogeneous and somewhat intractable character of the Web" (p. 310). Consequently, there are texts on the Web for which a successful classification is yet to be achieved. This is neither a trivial task nor a secondary one, since there are corpus linguists who believe that “text category is the most important organizing principle of most modern corpora"(O’Keeffe & McCarthy, 2010, p. 241). Renouf (2007) also claims that lack of metadata makes an exhaustive study impossible or at least undermines it. Potential register- or variety-based studies, which require a precise idea of production conditions and text genre, are a good example.

Corresponding to the potential lack of information concerning the metadata of the texts is a lack of information regarding the content, which has to be recorded and evaluated a posteriori.

> “Automated methods of corpus construction allow for limited control over the contents that end up in the final corpus. The actual corpus composition needs therefore to be investigated through post hoc evaluation methods." (Baroni et al., 2009, p. 217)

Thus, the problem is twofold, on the one hand it is a meta-information and categorization issue (Bergh & Zanchetta, 2008, p. 325), and on the other hand the actual contents of a web corpus can only be listed with certainty once the corpus is complete.

The issue of text typology General text typology is summarized above p. 13. Among the criteria mentioned by Atkins et al. (1992, p. 17), the following four are particularly problematic.

First of all, the makeup of a text is supposed to be “self-evident" (“a single text by one author is single"). But authorship of unknown web texts is much more difficult to determine. Second, the factuality criterion, which leaves “many problem areas" in traditional corpora (although literary texts in reference corpora for instance are not a problem), causes even more problems in web corpora, where the credibility of web pages and the value of statements in texts cannot be precisely assessed. Third, the setting of a text, i.e. “in what social context does the text belong?", may not be determined in an unequivocal manner due to evolving contexts and possibly unknown categories. The function of texts was “not easy" to assess to begin with according to Lemnitzer and Zinsmeister (2010) and it has become even harder, for instance because of new text types such as microtexts.

Last but not least, two criteria impact all previous ones, that being the text and language status. The first is supposed to take the values “original", “reprint", “updated", “revised", etc., which is not only difficult to retrace on the internet but also different in nature, for instance because of retweets, reblogs, or reposts. This category actually leads to an ubiquitous issue, the need for near-duplicates removal. Last, the language status (“source" or “translation") leads to complex text classification issues such as machine translation detection, and the identification of age and first language of text producers.

The notion of genre is everything but unanimously defined, all the more since new or unknown genres are bound to emerge over again in the context of web texts. The notions of authorship, mode, audience, aim and domain (Sharoff, 2004, 2006) are attempts to address this issue, as well as those of audience, authorship and artifact (Warschauer & Grimes, 2007). The difficulty to define new genres stable in time is precisely the starting point of a current research project involving Douglas Biber and Mark Davies (*A linguistic taxonomy of English web registers*, 2012-2015).

Specific issues are also to be found, for instance in the case of computer-mediated communication (CMC), with a growing range of genres as well as a rapidly evolving media universe. Internet-based communication possibly has to be tackled with ad-hoc tools and corpora, so that an ongoing project to build a reference corpus for CMC in German (Beißwenger, Ermakova, Geyken, Lemnitzer, & Storrer, 2013) involves a series of decisions and cannot be established from ready-made procedures.

Text categories as extrinsic and intrinsic composition criteria, a possible caveat of "traditional" corpora?

> "The design of most corpora is based on such external criteria, that is, using situational determinants rather than linguistic characteristics, as the parameters of composition." (Hunston, 2008, p. 156)

Additionally, several typologies can be articulated on text level (Habert et al., 1997), such as genres and registers on the one hand, and intuitive categories used by speakers that may evolve as well as invariant "situational parameters" on the other.

> "The text categories sampled in the Brown corpus have often been referred to as 'text types' or 'genres'. In the narrower, text linguistic sense, the use of this terminology is hardly justified. The categories are only a fairly rough-and-ready classification of texts. Research by Biber (1998) has shown, for instance, that sometimes more variation within traditional text categories (such as 'newspapers') exists than between different text categories." (Hundt, 2008, p. 171)

Press text is seen as close to the norm. It is supposed to replicate developments in society with relatively close coverage, allowing a user to look for new words or new expressions as well as to witness the disappearance of others.

However, as Hundt (2008) put it, "the question is whether one year's worth of *The Guardian* or *The Times* can be considered a single-register corpus or not" (p. 179). In fact, apart from the particular style of a newspaper, texts published online are increasingly a blend of locally produced texts and material from other sources.

Thus, the lack of control and/or precision regarding metadata is not necessarily typical for web corpora. The latter only make an already existing issue more salient.

> "It is worth pointing out that the lack of metadata is not unique to web corpora. Consider, for instance, corpora containing mostly newspaper articles (like the German DeReKo), where authorship cannot always be attributed to specific individuals."(Biemann et al., 2013, p. 47)

The lack of metadata is also a possible issue in the case of the BNC (see p. 27).

According to Kilgarriff and Grefenstette (2003), text typology is a case where there is still much to define, so that the Web as a potential resource even "forces the issue"[49]. Thus, it could be considered a telltale sign.

All in all, corpus categories are wished for, but so far they have essentially been a finite, well-known series of different types. Texts on the internet reflect a manifold reality which is

[49] "'Text type' is an area in which our understanding is, as yet, very limited. Although further work is required irrespective of the Web, the use of the Web forces the issue." (Kilgarriff & Grefenstette, 2003, p. 343)

difficult to grasp. Some see it as a downside of web corpora that one has to cope with less meta-information and with more a posteriori evaluation of the content. However, one could also say that these two characteristics exemplify tendencies that were already present in traditional corpora, as well as issues that were not properly settled, including the very operative definition of genres or registers.

1.4.2.3 Representativeness of web texts

Obviously, texts taken from the Web do not constitute a balanced corpus in a traditional sense, mostly because if nothing is done in order to establish such a balance, then textual material is not controlled, limited or balanced in any way (Bergh & Zanchetta, 2008, p. 325). The issue of representativeness as it is understood in corpus linguistics is discussed below. For a discussion of web representativeness, see p. 126.

Representativeness and typology The issue of representativeness follows directly from the potential lack of meta-information described above. Be it on a corpus design level or on a statistical level, it is impossible to know to what extent the gathered texts are representative of the whole Web, first because not much is known about the texts, and secondly because the composition of the Web is completely unknown, and can evolve very quickly in the course of time or according to a change of parameters, such as language or text type.

> "It is rather complicated to do [...] stratified sampling with web data, because (i) the relative sizes of the strata in the population are not known, and (ii) it would be required to start with a crawled data set from which the corpus strata are sampled, as web documents are not archived and pre-classified like many traditional sources of text. Web documents have to be discovered through the crawling process, and cannot be taken from the shelves." (Biemann et al., 2013, p. 24)

Precisely because a general approach to web corpora involves a constant discovery of web pages and web documents, the result cannot be known in advance and thus cannot be balanced. Nonetheless, it does not mean that the discussion about representativeness and balance does not apply to web corpora. It may indicate, however, that this discussion will not yield satisfying results.

Redefining the issue The representativeness issue may be typical of the corpus linguistics community, but it does not seem to be very intelligible, particularly to web scientists:

> "There are arguments about how representative the Web is as a corpus, but the notion of what a corpus should represent – should it include speech, writing, background language such as mumbling or talking in one's sleep, or errors for example? – is hard to pin down with any precision." (Berners-Lee et al., 2006, p. 50)

It also seems to show how corpora from the web exemplify existing issues and debates in the field of linguistics, which may have been ill-defined and as such need to be discussed:

> "The Web is not representative of anything else. But neither are other corpora, in any well-understood sense. Picking away at the question merely exposes how

> primitive our understanding of the topic is and leads inexorably to larger and altogether more interesting questions about the nature of language, and how it might be modeled. (Kilgarriff & Grefenstette, 2003, p. 343)

One may also say that it is cumbersome to pay too much attention to a notion that is not precise or adapted enough in a web context. Regarding this, the idea that corpus size may solve conceptual problems can also be discovered by linguistics, in this case by lexicographers:

> "In a billion-word corpus, the occasional oddball text will not compromise the overall picture, so we now simply aim to ensure that the major text-types are all well represented in our corpus. The arguments about 'representativeness', in other words, have lost some of their force in the brave new world of mega-corpora." (Rundell, 2008, p. 26)

A potential way to cope with representativeness would then be to aim for global composition requirements loose enough not to get in the way of corpus building. A "weak" understanding of representativeness seems indeed to pave the way to compromise.

Possible solutions On the corpus linguistics front, Leech (2006) developed a reception-based estimation of representativeness. However, other researchers such as Atkins et al. (1992) are in favor of a balanced ratio of production and reception:

> "The corpus builder has to remain aware of the reception and production aspects, and though texts which have a wide reception are by definition easier to come by, if the corpus is to be a true reflection of native speaker usage, then every effort must be made to include as much production material as possible." (Atkins et al., 1992, p. 7)

Web corpora make such a balance possible, precisely because or in spite of web page interlinking biases as well as website audience statistics. However, the tradition is not to worry about sampling, as long as there is a certain degree of variation. (Schäfer & Bildhauer, 2013, p. 31)

> "[Results suggest that] Web corpora built by a single researcher literally in minutes are, in terms of variety of genres, topics and lexicon represented, closer to traditional 'balanced' corpora such as the BNC than to mono-source corpora, such as newswire-based corpora." (Baroni & Ueyama, 2006, p. 32)

Maybe because of biases in the way texts are collected, and prominent pages being favored in the process (for a comparison of sources see chapter 4), the results can be considered to be acceptable, especially with respect to traditional reference corpora.

1.4.2.4 Suitable toolchains and processing speed: practical issues

A user-friendliness problem First of all, one may say that corpora aiming for web scale have a problem with "user-friendliness (as the Web was not originally designed for linguistic research)" (Bergh & Zanchetta, 2008, p. 325). It affects both the corpus builders and the users.

In fact, the web as a corpus framework was fashionable in 2004 with the launch of the WAC-workshop. Then a few major contributors left, and it began to get more and more complicated to gather corpora as the web kept expanding and diversifying, e.g. with the Web 2.0 and social media.

Moreover, the weaknesses of generic formats also account for some difficulties. Text encoding schemes, such as XML TEI, are not easy to apply to web texts, since they are primarily conceived for printed texts. Standards in general are not easy to adapt to the new reality, for instance dating systems. There are several ways to date a text, on the one hand on a text-to-text basis, with the time of last modification of the file on the web server, the redaction time as advertised in the content or in the metadata, or the creation time of the page, and on the other hand on a corpus basis, with for instance the time of retrieval by the corpus builders, or the first release of the corpus.

Similarly to the standards issue, scalable retrieval and querying infrastructure may impede adoption of web resources by linguists used to well-documented corpora and tools as well as to stabilized architectures. Most corpus construction projects are works in progress, so that software and content updates can be frequent, and results as well as display are usually constantly improved, which can affect the global user experience.

On the other side, there are also early adopters of search engines such as Google who are used to the apparent simplicity of the interface and who may be confused by the profusion of details of query syntax, subcorpora, or simply information overflow. This may explain why there are linguists who still try their luck on search engines directly (see the remarks on Googleology p. 43).

Too high a cost? Processing speed should not be as much of a problem as it was in the 2000s for web corpora or even in the 1980s for digital corpora. However, in a context of expanding web size and decreasing expenditures on public research, the situation is not favorable to researchers.

Tanguy (2013) states that it might be too costly for a research institution to address the web as a whole[50], because the material costs for running a crawler and extracting text are much too high for academic budgets.

In recent articles on web corpus construction, no one claims to indeed truly harvest data on a web scale, in the sense that research teams compete with commercial search engines in terms of computational power or pages seen and actively maintained. The adjective “web-scale" typically refers to large corpora, meaning that they could not be from other sources than the web, but it does not mean that the corpora are truly on the scale of the web, or even one or two orders of magnitude smaller.

However, the claim by Tanguy (2013) should be kept in perspective. With the profusion of open-source software most tools are already available and need not be specially crafted for a particular case. Thus, it is not necessary to invest much time and energy in document retrieval, but “merely" computing power, which becomes cheaper over the course of time.

That said, corpus processing is tricky, as shown in chapter 3. All in all, I prefer to see Tanguy’s claim as a call for light-weight, more efficient approaches, for example by restricting

[50]La création d’un moteur de recherche, ou plutôt d’un crawler capable de parcourir le Web pour en extraire le contenu textuel est un travail de très longue haleine, et le coût matériel de son fonctionnement est colossal, bien hors de portée des budgets académiques."(Tanguy, 2013, p. 14)

the field of inquiry, so that even with an expanding Web, it is still possible to grasp a significant part of it.

Linguistic processing: Normalization and annotation As a last point, normalization of web texts can also be a problem. There are always new and unexpected kinds of typographical errors as well as erratic or absent punctuation (Renouf, 2007), distorting linguistic results downstream, and yielding for instance deviant orders of magnitude for type/token ratios (see chapter 4 for a study).

Once confronted with this difficulty, it is possible to try and estimate how characteristic the type distribution is, but it is not easy to remedy the problem, as unduly correcting “mistakes" may alter the quality or the interest of the material.

Linguistic processing tools such as tokenizers and part-of-speech taggers are usually trained and tested on traditional reference corpora, i.e. on data sets which are drastically different from certain web corpora. This situation can lead to frequent data sparsity problems, or more generally unexpected automatic decisions.

Finally, the proper categorization and documentation of corpora relies on the good functioning of annotation processes (Lüdeling, 2006), meaning that an unexpected error rate at the beginning of the toolchain compromises further processes.

The adaptation of tools to the concrete reality of web corpora, i.e. of language as spoken on the Web, is certainly necessary to provide a suitable ecosystem for researchers. However, as it is part of corpus processing and not of corpus construction by itself, this topic is beyond the scope of this thesis.

1.4.2.5 Documentation

The nature of corpus documentation has changed dramatically, since web corpus building technology may not be accessible to linguists, as it is part of the software (if it is open-source), which linguists cannot be expected to know, or because it takes place on a technical level and requires a precise understanding of algorithms and web science.

According to Habert et al. (1997), a corpus is “stillborn" if it is not delivered with its documentation. They mention that due to the growing ease with which electronic resources are gathered, design decisions and research objectives can be easily forgotten, making the corpus unusable.

Modern web corpora do not satisfy these constraints, yet they are used by linguists, illustrating a shift in research practice which is mostly related to the growing complexity of web corpus construction. In fact, web crawling for example may require skills that are far from core linguistic knowledge and cannot be expected of linguists or even computational linguists. There might even be an expanding gap between digital humanities in general and web crawling, as the program of the 2014 edition of the general assembly of the International Internet Preservation Consortium[51] shows, with a workshop dedicated to “crawl engineers and operators" meant to “explore ways of performing more complex crawls to deal with specific problems", which certainly has to do with the growing complexity of one of the most used crawling frameworks, Heritrix.[52]

[51] http://netpreserve.org/general-assembly/2014/overview

[52] http://crawler.archive.org/

1.4.2.6 Republication, legal constraints and risks

The accumulation of manipulable data also bears risks from a legal and ethical perspective. There is progress to be made towards clear notions of ownership and privacy applied to the en masse availability of data which has been published separately. The potential conflicts between database and ownership rights in Germany are an example of a legal debate concerning the release of aggregated data sets.

Copyright issues Moreover, the existence of public domain texts is essential when it comes to transmissibility of corpora and results, because the question of ownership rights does not seem to be easily settled globally, making it all the more difficult for researchers. In my own experience, it is for instance much more difficult in Germany than in France, due to a more present and more restrictive legal framework and also due to a more consensual tradition of compliance with existing laws.

> "The copyright issue remains a thorny one: there is no easy way of determining whether the content of a particular page is copyrighted, nor is it feasible to ask millions of potential copyright holders for usage permission. However, our crawler does respect the download policies imposed by website administrators (i.e. the robots.txt file), and the WaCky website contains information on how to request the removal of specific documents from our corpora. Lastly, it must be noted that we offer highly processed versions of the web pages we download, in a format unlikely to be usable by non-linguists or for non-research purposes." (Baroni et al., 2009, p. 224)

For example, a study using newspaper articles, be it within the same source or across several ones, typically faces the kind of legal restrictions described above. It is possible to obtain articles simply by downloading web pages, it is possible to analyze them and to publish scientific results as well as text snippets or short quotes used as examples. However, making the annotated text available is not allowed, even if PoS-tagged text in an XML format clearly targets a rather narrow community.

All in all, there are corpora which are not designed to be a public park as it were but rather a "hortus conclusus", i.e. a walled, secluded garden. Newspaper articles, for example, are very frequent on the Internet, and some of them attract numerous readers. They may interest linguists for various reasons, however, they exemplify the discrepancy between ubiquitous content and impossibility to republish content other than in the form of smallish quotes.

Ethical questions If data are supposed to be in the public domain, they can be considered free of copyright concerns, but should not be treated as unworthy of questioning on an ethical level.

> "The process of evaluating the research ethics cannot be ignored simply because the data are seemingly public. Researchers must keep asking themselves – and their colleagues – about the ethics of their data collection, analysis, and publication." (boyd & Crawford, 2012, p. 672)

For instance, possible privacy issues can arise from intersections that are made in a large set, which, according to the metadata delivered with the corpus, can lead to identification of

individuals, their localization, or revelation of patterns in their daily lives. Thus, it should be ensured that such intersections cannot be used to harm anyone prior to corpus release.

“Masking" corpora is viable but not desirable Apparently, identification of facts and persons as well as reconstruction of whole texts seems to be a problem. This is where the necessity to use “masking" techniques (Rehm, Witt, Zinsmeister, & Dellert, 2007) comes from. Masking features for instance the replacement of tagged words by others which are randomly chosen within the same grammatical category, or the so-called “scrambling" of sentences prior to publication, where sentence order is randomized. Both methods make a reconstruction of the original text impossible but as such they also hinder a whole range of linguistic studies, mostly those basing on inter-phrasal or discourse level. Cohesion as well as coherence are lost, but lexical frequencies or infra-phrasal syntactic phenomena remain intact in the case of scrambling for instance, which is a start.

A more detailed description of the issues at stake can be found in my article on methodology (Barbaresi, 2011b). To sum up, from an open science perspective, it is crucial that scientific instruments and data are transmissible, in order to allow for reproducibility of experiments as well as a tighter collaboration between scientists. To this end, corpus construction from sources freed from copyright restrictions is mandatory, which implies being able to determine the license under which a given web text has been published. An example is given in chapter 4 using blog posts (see p. 221).

1.5 Intermediate conclusions

1.5.1 (Web) corpus linguistics: a discipline on the rise?

A new generation of researchers, corpora of a different nature There is a strong belief among corpus linguistics enthusiasts that the trend towards more empirical research in linguistics will continue, not least because of a generation change:

> “Max Planck was one of many who have pointed out that a new scientific approach does not win acceptance by convincing opponents, but by the fact that the opponents eventually die off and a new generation grows up familiar with it. At present, unempirical linguistics is still being written and read. But it seems safe to say that this is a temporary state of affairs." (Sampson, 2013, p. 288)

This change of generation is twofold: on the technological side, it brings more tech-savvy linguists, especially in computational linguistics, and on the conceptual side, it brings a rise of the search for regularities and patterns from a statistical view of language, with sometimes purely statistical methods, empowering a different perspective on language and corpora.

Bergh and Zanchetta (2008) speak of a “general cultural revolution caused by the emergence of the Web" (p. 310), but it would be simplistic to credit the Web alone with paradigm changes that are much deeper and which concern a whole range of disciplines. The technosciences and big data are two examples of new paradigms, discourses, and frameworks, affecting the way science is made, among other things through a shift in research policies and funding:

> “Due to various political and economic changes around the world, there is a greater emphasis these days on deliverables and evaluation. Data collection efforts have

been relatively successful in responding to these pressures by delivering massive quantities of data." (Church & Mercer, 1993, p. 21)

It is important to see that linguistics is not the only discipline in humanities to be concerned by these changes. The existence of data-intensive quantitative methods also confronts sociology, for example, with methodological breakthroughs or doubts:

> "'Numbers, numbers, numbers', writes Latour (2009). 'Sociology has been obsessed by the goal of becoming a quantitative science'. Sociology has never reached this goal, in Latour's view, because of where it draws the line between what is and is not quantifiable knowledge in the social domain." (boyd & Crawford, 2012, p. 666)

Concerning the notion of quantitative science, one may say that statistical approaches in linguistics have had a similar effect. The differences in research culture, notions, and experience leave a scattered field where opinions still diverge.

Enthuasiasts and skeptics After all, there are corpus enthusiasts, but also skeptics. The first are convinced of web corpora, in which they see the opportunity to observe linguistic phenomena and find evidence that, due to the increasing influence of corpora, cannot be refuted in their eyes:

> "Where a given usage can be shown to be both frequent and widespread, there is no questioning its status as being 'in the language' and therefore worth recording." (Rundell, 2008, p. 27)

For this reason, some consider web corpora to provide a "great momentum for further advance in the field of corpus linguistics" (Bergh & Zanchetta, 2008, p. 310).

On the other side, the skeptics argue that the opportunistic approach behind larger corpora is questionable, since a whole tradition is ignored in favor of texts which seem to have been included in corpora simply because they were available:

> "Compiling a proper text corpus entails a much greater amount of work than merely collecting any kind of text that you can lay your hands on, especially where other text types than newstext are difficult or impossible to acquire in electronic form." (Borin, 2009b, p. 6)

Nevertheless, precisely because of its open understanding of the notion of corpus, the current of "web linguistics" (Bergh & Zanchetta, 2008) exemplifies a number of unquestioned aspects of practice in corpus linguistics .

Corpus linguistics to its end : similarities and discontinuities The "magnifying glass" (Hundt et al., 2007) formed by web corpora may be considered by part of the research community as a change of paradigm towards statistical methods and "quantitative" linguistics, a framework adapted to larger corpora making it possible to "sift through" "heavily-diluted" data (Atkins et al., 1992).

In fact, the notion of serendipity may play a greater role than is admitted or considered acceptable by the linguistic tradition, be it in manual and semi-automatic research of attested

data, or in data-intensive use of corpora. In the latter case, it is expected that knowledge in a statistical sense may be extracted from corpora, for example in the form of language patterns determined or studied as a direct function of corpus coverage, which advances it to an important characteristic.

The magnifying glass also exposes the blend of research goals and processing issues in research practice, with “a ‘corpus architecture’ that is taken for granted" (Mair, 2012), partly because of epistemological gaps in corpus linguistics but mainly because there is too much to observe and too much to do, so that applicative scenarios are preferred.

The traditional criteria of sampling, representativity, and authenticity as they were considered in classical corpus construction are simply not applicable to general-purpose web corpora. However, they might still be productive when it comes to specific corpora, and it does not mean that they have lost their relevance. The change of perspective implies a research methodology that still has to be defined and a consensus on web data which still has to be found. Internet corpora can either ground on existing standards or they can complement or replace them.

The “opportunistic" way of gathering linguistic data from the web has already convinced those researchers who are now its supporters. The materials are there, they are available, and processing tools are available too. But there are still efforts to be made in order to reach a larger spectrum of users, gain more traction from others, and simply be fully accepted as a valid data source. That is why a minimum of critical examination as well as a nod towards a tradition of qualitative text scrutiny should be preliminary to a study.

1.5.2 Changes in corpus design and construction

Web corpora are an unprecedented opportunity for observing linguistic phenomena rarely seen in written corpora and to apprehend new text types which open new issues regarding linguistic standards and norms, contact and diversity. Web texts may belong to new genres, resulting from practices developed at the age of internet-based communication. Statistically speaking they may also give a broader access to text production by a large number of existing socio- and ethnolinguistic groups, including lesser-known ones.

A matter of perspective: reopening the case of word order in German The word-order example, first mentioned in the introduction on p. 6, makes it clear that according to the perspective given by a corpus, divergent conclusions on language can be drawn. For instance, grammarians working on reference corpora including mostly “traditional" written text genre such as newspaper articles and novels, may rule out the rare cases of verb-second subordinates they encounter. Then they may conclude that the subordinate clause with the verb at the end is the standard structure in German. This is even more true if the sentences are hand-picked, because theoretical linguists may then have a tendency to choose “favorable" sentences with respect to the theory they are trying to prove.

In contrast, web corpora containing more casual genres or even speech transcriptions such as subtitles or spontaneous blog comments may lead to other conclusions. First, despite the meaning of “representative" in traditional corpus building, they may present a more representative image of how language is currently spoken by a broader spectrum of speakers. In that case, the relative abundance of verb-second subordinates implies that they are bound to be detected. Second, the use of more quantitative methods may also lead to the conclusion that in spite of numerous subordinate clauses following the “classical" model, verbs in second position

are clearly present in a majority of cases, because there are more principal clauses than subordinate clauses and because the verbs do not always come at the end in the latter case. That is why German is generally considered to be a V2 or a flexible language in language typology. For instance, it is classified by the World Atlas of Language Structures (Dryer & Haspelmath, 2013) as having “no dominant order".

One may argue that the sheer number of occurrences of a given structure does not necessarily illustrate its importance. That is why theories in linguistic research are often a matter of perspective. Since corpora play a major role in empirical linguistics, their origin and their composition ought to be better known.

Summary: characteristics of “pre-web" corpus construction

- Typology fixed most of the time before construction: find texts for all classes and balance the whole
- Normalization: sometimes before, sometimes after. Considered to be important.
- Complete texts and/or extracts and/or derivates (Ngrams, word frequencies)
- Persistent vs. temporary corpora
- Metadata: rich, sometimes manually edited or poor/not verified
- Classification according to text production parameters or according to tools output (machine learning, statistical criteria: clustering)
- Subcorpora: used for balancing or register comparison

Because they are often difficult to access and difficult to process, non-standard variants are usually not part of reference corpora.

Post-web changes The prototypical “web as corpus" construction method is an extreme case where the corpus design steps mentioned by Atkins et al. (1992) (see p. 11) are reduced to the last two: data capture followed by corpus processing.

The shifts listed below are changes as compared to “pre-web" corpora:

- Be it for precise target and typology for specialized corpora and focused crawling, or concerning exploratory corpus construction for general-purpose web corpora, it is necessary to consider texts, text types, and text genres beyond the previous extension of these notions and beyond known categories.
- The normalization of the documents is much more delicate, it can be done *a posteriori* as a particular step or be left out, as it may be considered as error-prone.
- Usually, web corpus construction deals with the retrieval of complete texts, but due to the nature of HTML documents and the necessary post-processing (see the following chapter), they may be constructed of extracts that are artifacts of the text type and processing tools, and thus differ from the traditional sense of “extract".
- The classification becomes a major issue.

- Consequently, there are usually no subcorpora as such and general-purpose corpora are taken as a whole, while specialized corpora are divided into categories as long as available metadata allow for such an operation. However, web corpora also lead to an abundant production of derivates: n-grams, word frequencies, language models, training sets for diverse tools.

By contrast, the perspective stays the same concerning the following topics:

- There are persistent and temporary corpora. A given corpus may be extended regularly, and thus correspond to the notion of monitor corpus.

- The metadata may be rich for some specialized corpora, but poor otherwise.

The following facts about digital corpora are not necessarily true anymore regarding corpora from the Web:

- "All corpora are a compromise between what is desirable, that is, what the corpus designer has planned, and what is possible." (Hunston, 2008, p. 156)
  In fact, general-purpose web corpora are rather a compromise on the technical side. Since there is no general web cartography and often no design plan, it is not possible to assess the recall of a web corpus with respect to a population of web documents.

- "Any corpus, unless it is unusually specific in content, may be perceived as a collection of sub-corpora, each one of which is relatively homogeneous." (Hunston, 2008, p. 154)
  The homogeneity of these subparts is not guaranteed anymore, nor are corpora really perceived as such a collection. When this is the case, the characteristics shared by web texts are of a different nature than the criteria used to build subcorpora, and most of the time such common characteristics must be inferred from the data and cannot be determined *a priori*.

- The most important practical constraints are software limitations, copyright and ethical issues, and text availability. (Hunston, 2008, p. 157)
  Software limitations can now be considered secondary, as well as text availability, since it is the very condition of text inclusion. Ethical issues and copyright are not really a constraint, but rather a factor.

1.5.3 Challenges addressed in the following sections

From the viewpoint of corpus users, the uncontrollability of the data appears to be a principal caveat, as it appears to be impossible to make design decisions regarding the nature, type, or quality of the texts that become part of the corpus. On the side of corpus builders, unquestioned data collection and data querying processes leave room for improvement, while the relatively new field of language documentation draws special attention to the recording, processing and preserving of linguistic data.

Concerns regarding data quality and exploitability in linguistics The unquestioned and for some linguists ill-founded way of dealing with web data leads to a great deal of skepticism towards what is described as opportunism (Joseph, 2004).

The Web is not only text, rather the text has to be filtered out, which it neither an obvious nor a lossless process (cf chapter 3). And when it is text, the Web generally leaves "a lot of noise, such as automatically generated non-linguistic material and duplicated documents" (Baroni & Ueyama, 2006, p. 32), two issues which are tackled in the next chapters.

As will be explained further, the question whether the method and tools used so far in web corpus construction provide a good overview of a language is still open. In most cases, despite or because of its convenience, it is not proper corpus construction, i.e. a mastered process responding to essential research questions and following a precise research methodology.

In order to address these concerns, it seems necessary to work on the text criterion, and to find and describe a norm, and/or use a source of linguistic material about which there is a broad consensus regarding its reliability. Both issues are tackled in chapter 4.

Accessible, independent, and practically doable web corpus gathering I agree with Baroni and Ueyama (2006) as they discard approaches relying directly on search engine data and as they strive to make linguists independent from them:

> "We believe that the only viable long term approach to constructing Web corpora is for linguists to perform their own crawls of the Internet. This makes linguists fully independent from commercial search engines, and provides full control over the whole corpus construction procedure. However this is also the most difficult approach to implement, especially if the target is a large corpus." (Baroni & Ueyama, 2006, p. 33)

However, the authors note that in order to do so, considerable computational resources are necessary, an issue also mentioned by Tanguy (2013). Moreover, research data need to be cleaned and annotated (Baroni & Ueyama, 2006, p. 33).

Meeting diverging expectations halfway All in all, a possible goal would be to try to take the best of both worlds: the "big is beautiful" axiom on the one hand, and on the other the idea that "carefully selected research data trump uncontrolled datasets", by offering concrete solutions somewhere in between.

The present work can also be considered as an attempt to satisfy multiple constraints symptomatic for particular traditions and practices, for example the French appetite for abstraction and synthesis on one hand, with the roundup and discussion on corpus linguistics in this chapter, and the empirical trend towards data and modeling coming from the English-speaking world and in turn potentially more present in Germany than in France.

Chapter 2

Gauging and quality assessment of text collections: methodological insights on (web) text processing and classification

2.1 Introduction

Text qualification In the previous chapter, a history of corpus building was sketched, where I showed to what extent web corpora differ from traditional ones. In short – and in terms of garden design (see p. 2) – I showed how to put together components for a garden as well as how to arrange them. This chapter instead is more about providing an overview to see if the garden fits its design objectives, as well as trimming the trees and separating the lettuce from the nettles.

One of the more salient characteristica of web texts is that they make it easier to gather corpora, be it general-purpose or specialized ones, since there are more texts available than ever before. However, it is also more difficult to assess their validity with respect to research objectives (see p. 56).

Additionally, the diversity of potential corpus usage scenarios poses questions concerning good practices in addition to the usual problems and solutions. More specifically, web texts raise issues concerning a wide range of text cleaning and mining procedures, for instance in terms of robustness, manageability and adequacy. Web corpus preprocessing issues are summarized in the next chapter.

More generally, text qualification, i.e. gauging and adjusting existing collections, does not only concern corpus linguistics, but also other disciplines. It is relevant in that sense to examine how problems are addressed and eventually solved in different communities. Sometimes the approaches are very different, even incompatible, and sometimes there are common features to be found.

Outline Text quality assessment is a recent field of inquiry which deals with underestimated but characteristic problems of web corpora. It is partly a binary classification problem with respect to the texts or parts to discard or maintain in the corpus.

Research on readability is an active topic, usually not seen as having a binary but rather a multi-class output. It features industrial interest with approaches firmly in application as well as different trends in linguistics, psycholinguistics, and natural language processing.

Last, corpus visualization has seen various approaches, first from computer science and more recently from digital humanities. All in all, it is still at an early stage, but seems to be a promising way to get to know a corpus, explore it, and discover possible flaws.

2.2 Text quality assessment, an example of interdisciplinary research on web texts

2.2.1 Underestimated flaws and recent advances in discovering them

2.2.1.1 Introduction to text quality issues

Text quality variation throughout the corpus can be a consequence of corpus design, of document collection processes, and/or preprocessing. For a number of reasons described in the paragraphs below, and despite ongoing work on the issue, there are many different subsequent questions to address. They concern text quality, entailing much work, and more precisely further interdisciplinary work.

In fact, quality does not only fall within the field of computational linguistics. Other disciplines such as information retrieval have seen a similar evolution, where problems afferent to text quantity such as document gathering on a large scale as well as scalability of procedures can be considered (at least) as partly solved. Attention is now turned towards text quality.

In information retrieval particularly, there seems to be a void concerning web document quality analysis:

> "Many existing retrieval approaches do not take into account the content quality of the retrieved documents, although link-based measures such as PageRank are commonly used as a form of document prior." (Bendersky, Croft, & Diao, 2011)

Because of particular affinities for graph-based approaches in computer science, link-based indicators have been favored as a matter of fact. But even in the case of documents returned as they are to the end user, i.e. without annotation or linguistic processing, discrimination based on quality metrics are gaining in popularity.

The question whether bigger data is better data where it concerns corpus linguistics is discussed in the previous chapter (see p. 35). The remainder of this section turns to specific examples of text quality issues in natural language processing and beyond.

2.2.1.2 Examples

In order to get a glimpse of the phenomena involved in text quality assessment, a few examples of the difficulties are listed below. They are grouped into four different kind of issues: automatically generated text, machine-translated text, human-induced spam, and multiple languages. The list is not supposed to be exhaustive, it rather summarizes a few salient problems for which studies have been undertaken. The tasks are described in section 2.2.1.3, while the current approaches in several research fields are evoked in section 2.2.2.

Automatically generated text

The following kind of repeated text[1] is probably induced by the crawler, which because of its technical specifications hinders a completely dynamic rendering of a given web page. A human user using a state-of-the-art browser would possibly see normal text injected from another source at this particular point of the page:

```
There was an error sending the message, please try again later.
description not available ...
There was an error sending the message, please try again later.
```

It is a basic but frequent kind of machine-generated text. Such sentences or paragraphs very often contain no useful information, and because of their repetitive nature, all the more on a scale of a whole website, there are clearly unwanted in the case of a corpus for linguistic studies.

Machine-translated text

Machine-translated text is sometimes hard to distinguish from a low-level of proficiency

[1] Extracted from test data analyzed in (Schäfer, Barbaresi, & Bildhauer, 2013)
http://www.carocean.co.uk/for-sale-Renault+Colchester.html

by a speaker or even from automatic content templates. However, the example below makes clear how different it is from normal utterances in a target language. On one hand, here is an occurrence of what is called by Arase and Zhou (2013) the phrase salad phenomenon:

> Of surprise was up foreigners flocked overseas as well, they publicized not only Japan, saw an article from the news.

On the other hand, here is how the authors translate it into natural English:

> "The news was broadcasted not only in Japan but also overseas, and it surprised foreigners who read the article."[2]

Since the example concerns the pair Japanese-English, which is notably more difficult to translate than translations within the same language family for instance, the "falseness" of the result is particularly striking. It is no rare case however, since for various reasons machine-translated content now belongs to the very essence of a user experience (see p. 76 for a discussion).

Human-induced spam

Real spam is sometimes more tricky to identify than machine-generated or machine-translated content, of which it can be a subcategory, as it is mostly the result of a human intervention. One of the most salient cases of spam is found in vague and elusive blog comments, which fit all cases and thus can be posted nearly everywhere, in order to advertise a product or simply to point links to a certain website, as was probably the case in these two examples[3]:

- This is certainly a amazing article. Thanks a lot for making the effort to explain this all out for us. It is a great help!
- It is a fantastic post. I will be so thrilled the web is still equipped with wonderful content material.

The potential damage in terms of corpus occurrences is lower than in the other types of issues. Nonetheless, a web-scale corpus is bound to gather numerous examples of such sentences, which efficiently distort the view on language offered by the corpus.

Multiple languages

The following three comments[4] were found just one after the other on a single web page. They are most probably blog comments following the same principles as human-induced spam mentioned just above.

> Das GmbH-Haus steht Ihnen sowohl für die bürokratische Abwicklung von Ihrem Gesellschaftskauf oder Unternehmenskauf zur Verfügung als auch bei der Etablierung verschiedener Firmenmäntel. Jeder Firmenmantel

[2] Source: (Arase & Zhou, 2013, p. 1599), see below for a more detailed analysis

[3] Both comments extracted from test data analyzed in (Schäfer et al., 2013)

[4] Extracted from test data analyzed in (Schäfer et al., 2013)
http://learnanatomyandphysiology.co.uk/about-2?replytocom=23533

(z. B. der GmbH Mantel oder der AG Mantel) birgt unterschiedliche steuerliche Vorteile. Gerne erläutern wir Ihnen die steuerlichen Möglichkeiten bei einem Mantelkauf.

I'm thus successful to own think of this site. An individual practically declared me just what My partner and i opted to be able to take note to be able to and also afterward a lot of. Amazing publishing and also all the best again regarding accomplishing the following simply no fee!

Lavoro per il domani - uno sguardo di Yesturdays ad alcuni esempi

Obviously, they are written in three different languages (respectively German, English and Italian). In fact, each paragraph taken apart is perfectly sound, although the English and Italian examples are not as structured as the German one, but the presence of all three on the same web page typically lowers one's expectations regarding content quality. Even without taking a closer look at them one may think they are most probably spam, and as such text parts that ought to be deleted.

However, this last type of difficulty is trickier, since its perception depends on the filtering level. On sentence or even paragraph level there is no problem at all, parts which do not correspond to the target language can be left out quite easily. On the web page level one does have a problem, since a series of heuristics have to be applied in order to systematically decide whether to keep the page or not. In such cases, the amount of text in the target language would probably be a productive criterium.

The issue of mixed-language documents is discussed in the next section, while experimental results on web document selection are treated further below (see chapter 4).

2.2.1.3 Discussion: Machine-generated/translated content, traps, and spam

Machine-generated content and traps Apart from the cases exemplified just above, machine-generated content may also serve the purpose of tricking other machines, especially crawlers, into falsely assessing a web page's content or content quality, particularly for so-called "black hat" -i.e. potentially malicious- search engine optimization techniques. Deception mechanisms targeted at machines are called crawler traps. A frequent goal is to trick search engines into assigning a web page a higher rank than would otherwise have been the case, thus generating more clicks and more revenue.

> "Another phenomenon that inflates the corpus without adding utility is crawler traps: Web sites that populate a large, possibly infinite URL space on that site with mechanically generated content. [...] Not much research has been published on algorithms or heuristics for detecting crawler traps directly." (Olston & Najork, 2010, p. 226-227)

Since web data is harvested automatically, one's software can be expected to fall for crawler traps, which means that text filtering has to take it into account. Since crawler traps cannot be actively detected, like Olston and Najork (2010) state, their precise impact on the course of corpus construction and/or on the final content of a document collection is unknown. Together

with machine-generated content, crawler traps call for exhaustive filtering steps aiming at detection of duplicate content.

Machine-translated content The amount of machine-translated content on the Web varies by language. For high-density languages such as English, Japanese, and German, only a small percentage of web pages are generated by machine-translation systems.

According to Rarrick, Quirk, and Lewis (2011), among pages for which they identified a parallel document, at least 15% of the sentence pairs annotated for both English-German and English-Japanese appear to contain disfluent or inadequate translations. Still according to Rarrick et al. (2011), the amount of machine-translated content on the Web rises sharply for lower density languages such as Latvian, Lithuanian and Romanian. Latvian and Lithuanian had the highest percentages, with each over 50%.

The problem with this proportion of machine-translated content is twofold. On the one hand, it affects corpus construction directly because what linguists are after is content produced by real human speakers and not by machines. On the other hand, machine-generated texts do not appear to be flawless, which can impede corpus research at any level, thus requiring detection and filtering of such content:

> "The quality of these machine-translated sentences is generally much lower than sentences generated by native speakers and professional translators. Therefore, a method to detect and filter such SMT results is desired to best make use of Web-mined data." (Arase & Zhou, 2013, p. 1597)

This makes the case of low-density languages even more complicated, in addition to these languages already needing special procedures (see p. 54).

All in all, machine-translated content is a major issue, as is text quality in general, especially when it comes to web texts (Arase & Zhou, 2013). Detection of machine-translated content has been proven to be efficient when developing a machine-translation system, so that it cannot be said in this particular case that "more data is better data" (see p. 35 for a discussion of this axiom):

> "Trained on our filtered corpus, our most successful MT system outperformed one trained on the full, unfiltered corpus, thus challenging the conventional wisdom in natural language processing that 'more data is better data'" (Rarrick et al., 2011, p. 1)

Mixed-language documents First of all, one may want to be sure that the text is mostly written in a given language. There are many Web documents that are mixed, first because the content comes from multiple sources, second because web servers adapt depending on geographic information, and last because there is globally a majority of speakers who use several languages on a regular basis and who may switch between them. In this respect, European countries where speakers focus on a single language are usually an exception.

Mixed-language documents slow down text gathering processes (King & Abney, 2013), which is another thing particularly true for lower-density languages:

> "We found that the majority of webpages that contain text in a minority language also contain text in other languages." (King & Abney, 2013, p. 1110)

The characteristic the authors describe seems to be inversely proportional to the popularity of a language, with a very high probability to find monolingual documents for the most frequently spoken languages:

> "If a language is not spoken widely enough, then there is little chance of finding any text in that language on the Web. Conversely if a language is too widely spoken, then it is difficult to find mixed-language pages for it." (King & Abney, 2013, p. 1112)

These findings give interesting insights on actual language use in everyday web experience, which is a potential interest of web corpora, and more generally data harvesting on a web scale, since it makes it possible to capture global trends which could before only be extrapolated from sample studies. The variations between widely spoken languages and others make the case of less-resourced languages (see p. 54) even more complicated.

Several forms of spam The main cause for spam are business models grounding on imitation, falsification, or generation of content. While scams and phishing are probably spam forms that are better known to most Internet users, from the point of a web crawler the most frequent form is related to search engine optimization techniques. In that case, web pages act as empty shells which are designed to generate "link juice" for others by building up a net of supposedly valid and influential websites. As long as the sites appear to be legitimate, their "good reputation" in terms of link-based algorithms can be monetized.

> "Web spam is motivated by the monetary value of achieving a prominent position in search-engine result pages." (Olston & Najork, 2010, p. 227)

Thus, the main problem with page ranking spam is not that these sites are numerous, but rather the fact that their content is just a mere addition of templates with little or no "real" text, i.e. no text which results from a natural utterance by a speaker.

The great majority of texts available on the web are natively digitized. As such they do not include potential digitalization flaws such as optical character recognition mistakes. Nevertheless, idiosyncratic biases such as spam and diverse optimization techniques mentioned in this section show that general web content cannot be expected to be flawless. Therefore, web texts need to be scrutinized and carefully filtered and cleaned.

2.2.2 Tackling the problems: General state of the art

2.2.2.1 Converging objectives from several research backgrounds

Text quality assessment is a broad and very active topic. It interests researchers from several disciplines (data mining, NLP, IE / IR) in many ways, since here the following research objectives converge: machine translation, readability assessment, language acquisition (essay rating), textual entailment, web document ranking, and web corpora.

There are cases where the researchers simply choose to tackle an open issue and to apply the methods of their field, and others where they really specialize in measuring, predicting or classifying the text quality.

An interesting point is that researchers who have a computer science background sometimes work with linguists, because a precise linguistic analysis may be an advantage in this

area, and because not all problems can be solved using the classical document ranking and other IE / IR methods. This is mainly due to the fast-paced evolution towards an always greater fluency and adaptability of machine-generated text: spam and other types of compiled or translated text become more and more credible and in a way meaningful.

2.2.2.2 Concrete examples in NLP and Computer Science

The following examples detail existing approaches, from broad indicators in a whole document collection to more specific ones. Insights from the general approach in Information Retrieval, Machine Translation detection, text classification, user profiling on social networks, and detection of text coherence in NLP show several methods which have been used successfully, some interdisciplinary and typical for a certain research field. The next subsection shows in more detail how problems can be addressed in web corpus construction.

General approach in Information Retrieval Bendersky et al. (2011) originally deal with information retrieval problems. However, they use various quality criteria and show the positive impact of a multidisciplinary method. They look for statistically significant indicators in a wide range of different research traditions:

> "Our experimental results show that QSDM [a quality-biased ranking method] consistently and significantly improves the retrieval performance of text-based and link-based retrieval methods that do not take into account the quality of the document content. Statistically significant improvements in retrieval performance were attained for both ClueWeb – a general web collection, in which our method was able to improve the retrieval effectiveness and to promote relevant Wikipedia pages even after an application of a standard spam filter – as well as for GOV2 – a specialized corpus, which contained documents of differing quality, but no explicit spam." (Bendersky et al., 2011)

At the end, their criteria tackle content analysis, readability and spam detection issues (including number of tokens, stop words, entropy, links). They managed to improve the rate of relevant pages both in the document collection and search results.

Readability criteria as well as statistical tests in order to find salient cues are mentioned below (see p. 94).

Machine Translation detection There is a potential synergy between text quality assessment and machine translation detection, because researchers in this field agree that machine-generated content ought to be ranked below human-written content. There is even hope that classifiers designed to align sentences and/or parallel corpora might be used on monolingual documents:

> "In addition to machine translation, MT detection also has potential application in search engine indexing. It may be desirable to rank machine-translated pages below human-written ones. While some adaptation would be necessary to apply the classifier to monolingual documents rather than parallel documents, we believe that our general approach is applicable." (Rarrick et al., 2011)

So far, the typical machine translation approach consists of considering parallel texts. Criteria mostly include measures on or below token level, such as character and token counts (with subsequent ratios between both sides of aligned texts), out-of-vocabulary tokens, script type characteristics (e.g. the proportion of Latin or Cyrillic script), and proportion of tokens which have a direct match on the other side (Rarrick et al., 2011).

However, when web documents are taken into account, metadata such as URLs are considered relevant. Rarrick et al. (2011) also include URL statistics in their study, based for example on domain or punctuation type, which is now considered to be seminal work.

Phrase salad phenomenon Arase and Zhou (2013) decide to tackle a particular aspect of machine-translated text, the "phrase salad phenomenon", which they define as follows:

> "Each phrase, a sequence of consecutive words, is fluent and grammatically correct; however, the fluency and grammar correctness are both poor in inter-phrases." (Arase & Zhou, 2013, p. 1598-99)

The example given by Arase and Zhou (2013) and already mentioned above illustrates how machine-generated text can be correct on phrase-level, represented by vertical bars below, even if unnatural phenomena on the inter-phrasal level seriously undermine the intelligibility of the whole sentence:

| Of surprise | was up | foreigners flocked | overseas |
as well, | they publicized not only | Japan, | saw an article
from the news. |

Their strategy was designed with web texts in mind. What makes it stand out in comparison with other approaches in the same research field is that it focused on monolingual text, and not on parallel or comparable corpora. Their aim was not finding suitable text to train machine translation systems, but to separate machine-generated text from the rest in a web-mining approach. As a matter of fact, the authors' affiliation is Microsoft Research Asia, meaning that web search algorithms are a probable application. This also explains why the authors use web texts as input.

The method is supposed to be computationally efficient and to fit large corpora:

> "We focus on the phrase salad phenomenon that is observed in existing SMT [Statistical Machine Translation] results and propose a set of computationally inexpensive features to effectively detect such machine-translated sentences from a large-scale Web-mined text." (Arase & Zhou, 2013, p. 1597)

The metrics used operate on phrase-level within a given sentence, as the approach consists of finding phrases whose internal syntactical characteristics might be in line with expectations but whose inter-phrase structure is problematic. In order to do so, three different features are used:

> "We define features to capture a phrase salad by examining local and distant phrases. These features evaluate (1) fluency, (2) grammaticality, and (3) completeness of non-contiguous phrases in a sentence." (Arase & Zhou, 2013, p. 1599)

The fluency feature is captured by language models, meaning that irregular inter-phrasal patterns are detected from a statistical point of view. The grammaticality feature is a more syntactical approach, it aims at finding inconsistencies regarding tense or voice, while the completeness addresses patterns such as “not only... but also".

All in all, the features used are mostly of a statistical nature: the fluency is estimated using language models, the grammaticality using models based on part-of-speech pattern regularities, while the so-called “gappy-phrase feature" first relies on human analysis that is then transferred to machine in the form of a systematical detection. As the authors mention, grammaticality is a well-known research field to rely on, with a relatively large number of research contributions.

Spam seen as a classification problem The problem of web spam mentioned above can be treated as a binary classification problem, where documents have to be separated into two different collections.

> “The problem of identifying web spam can be framed as a classification problem [...]. The main challenge is to identify features that are predictive of web spam and can thus be used as inputs to the classifier." (Olston & Najork, 2010, p. 227)

Approaches in computer science tend to use machine learning on large data sets, and many tasks qualify as classification problems, i.e. problems potentially solvable with acceptable precision by machine learning algorithms. A great amount of work is spent preparing the data so that such an algorithm can be applied.

Spam and profiling on social networks Sentiment Detection on social networks does not work without a considerable amount of filtering. If no filters are used the proportion of spam and machine-generated text is exceptionally high, too high for general web corpus standards.

In that particular case, the emphasis often lies on length of tweets and token-based statistics rather than on text quality per se with short messages (Benevenuto, Magno, Rodrigues, & Almeida, 2010). Fighting spam means categorizing, not only tweets but also users and user groups. In fact, profiling according to user metadata such as age (as given by the participants) and interaction level with others seems to yield interesting results (Benevenuto et al., 2010).

Detection of text coherence Text coherence and sense detection generally means being able to detect if a text follows a given direction, if it has a global meaning as well as an argumentation structure. Successful detection software does not primarily deal with machine-generated texts, it has found real application for educational purposes such as essay grading.

Research centers such as the Educational Testing Service are looking for ways to automatically score essays written by children for example. Even if they are potentially successful, these methods are meant to complement the work of teaching staff and not to replace it, mainly because of the fundamental differences between the ways humans and machines process texts:

> “Automated Essay Scoring systems do not actually read and understand essays as humans do. Whereas human raters may directly evaluate various intrinsic variables of interest, such as diction, fluency, and grammar, in order to produce an essay score, AES systems use approximations or possible correlates of these intrinsic variables." (Attali & Burstein, 2006)

The notion of approximation is interesting here, since the authors start from an applicative perspective and do not try to reconstruct the meaning of a text, for instance by using ontologies.

There is one major common feature to the research work mentioned here: it turns to linguistic patterns to accomplish the task. The study of word fields by Klebanov and Flor (2013) adopts the same point of view, and extracts "word association profiles" from a meaningful corpus in order to allow for a comparison with a standard. Word pairs selected according to their frequency are used to see if a text is both focused and imaginative. As a matter of fact, untalented redactors as well as machines lack originality.

> "We describe a new representation of the content vocabulary of a text we call word association profile that captures the proportions of highly associated, mildly associated, unassociated, and disassociated pairs of words that co-exist in the given text." (Klebanov & Flor, 2013)

In this approach, too common word pairs may signal a lack of originality, whereas words which do not belong with each other may signal a lack of focus and too wide a spread with the expected text topic.

Finally, there are also pattern-based approaches to syntax. While lexical items and linguistic patterns have been related since the 1980s at least, research on syntax using patterns shows that this global approach to language is becoming increasingly popular. The work of Louis and Nenkova (2012) for instance is an attempt at assessing text coherence using syntax patterns:

> "We introduce a model of coherence which captures the intentional discourse structure in text. Our work is based on the hypothesis that syntax provides a proxy for the communicative goal of a sentence and therefore the sequence of sentences in a coherent discourse should exhibit detectable structural patterns." (Louis & Nenkova, 2012)

In conclusion, detection of text coherence exemplifies that proxies have to be found for phenomena which strike the human eye but which machines fail to detect. This implies a series of approximations and seems to come along with increasing use of linguistic patterns in NLP.

However, due to their objectives, the detection procedures described above require a supervised training phase and are limited to a range of nonfiction and preferably argumentative texts.

2.2.2.3 Addressing text quality in web corpus construction

Quality assessment by way of corpus comparison is used by Biemann et al. (2013) for similar corpora, corpora in the same language, but also corpora containing different languages. The article is a joint work by several German research centers presenting a synthesis of how web corpora are collected and processed.

Their approach grounds on a series of directly computable indicators which can be used to detect potential systematical errors or biases in the corpora. The authors are mainly looking for extreme values:

> "Extreme values for certain statistics are possible indicators of problematic/noisy objects which require further inspection." (Biemann et al., 2013, p. 36)

These indicators include the distribution of word, sentence, or document lengths, the distributions of characters or n-grams, and potential overlapping with well-known “empirical laws of language such as Zipf’s Law" (Biemann et al., 2013, p. 36).

More specific indicators are used in the case of the Leipzig Corpora Collection. These indicators include (Biemann et al., 2013, p. 37):

- two crawl-related criteria (largest domains represented in the corpus and their size, and number of sources per time period),
- two character-based criteria (number of different characters used and character frequencies),
- four criteria on word-level (distribution of word length, most frequent words, longest words among the most frequent ones compared to the longest words in general, words ending in a capitalized stop word),
- and two sentence-based criteria (shortest and longest sentences, sentence length distribution).

The criteria I will detail in chapter 4 are related to these, for further analysis see below.

The analysis of results includes a pinch of visualization, since irregularities are expected to be clearly identifiable on a plot:

> “On closer examination, the peaks turned out to be the result of boilerplate material and near duplicates, which should have been removed." (Biemann et al., 2013, p. 37)

I experienced similar peaks during my experiments. As figure 2.1 shows, there are clear irregularities in web document collections which are easy to spot, for example by plotting the length of documents.

Figure 2.1 has been generated using web document statistics gathered during web exploration with my toolchain (FLUX), described more in detail below in chapter 4. Be it in characters or in words, at the beginning or at the end or processing, the length of documents gathered during crawls usually exhibit a skewed length distribution. In that particular case the abnormally high values probably reveal the existence of numerous duplicates.

These are rather “basic but efficient" indicators. While it can be enough to detect the crudest, most salient problems, it will probably fail to detect machine-generated text of a certain quality as well as documents that do not qualify as text because they lack internal cohesion.

To my best knowledge, there is no available study so far which explicitly focuses on web text qualification for web corpus construction. Comprehensive results on this topic, based on web pages manually annotated as to their suitability for corpus inclusion, and featuring the selection of criteria for web text filtering are introduced in chapter 4.

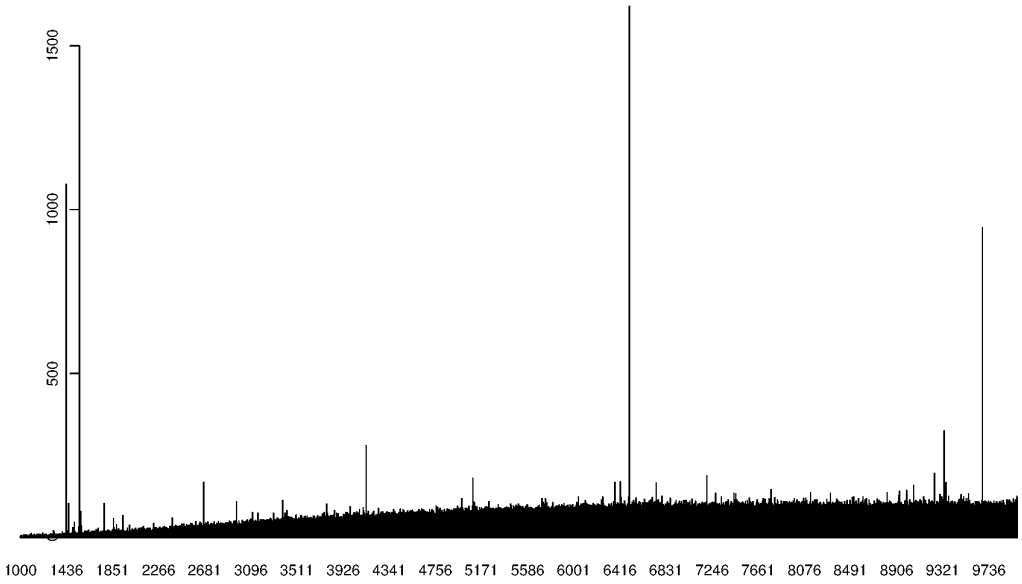


Figure 2.1: Length in characters before markup and boilerplate removal of documents in several languages gathered on the web, after basic filtering (zoomed view). Document length on the x- and number of documents on the y-axis. Peaks are clearly visible, although they are improbable, hinting at potential series of duplicate documents.

2.3 Text readability as an aggregate of salient text characteristics

2.3.1 From text quality to readability and back

Machine learning techniques are currently very popular in the research communities, not only in computer science but beyond that as well. One of the most notable areas here is NLP, as the second methodological example, readability, will show. It order to apply machine learning techniques efficiently, it is necessary to find the right textual clues, be it to fight spam or to assess the readability of a text.

Additionally, readability gives the opportunity to witness several methodological approaches as well as research trends in linguistics, psycholinguistics, and natural language processing: from “knowledge-poor" approaches relying on machine learning algorithms to expert approaches based on linguistic knowledge on the side of the researcher. This is the case for readability more than for text quality assessment, a discipline where linguistics do not historically play a role.

In that sense, the example of readability, comprehensibility or understandability studies is relevant to illustrate the approach used in text qualification.

2.3.2 Extent of the research field: readability, complexity, comprehensibility

Definition First of all, the notion of readability is very close to comprehensibility or understandability, though not in the general sense and not to the full extent of these words. It rather deals with a particular side, that being text difficulty:

> "Readability is a term commonly used to denote legibility of handwriting or typography, ease or pleasantness of reading, or to refer to the understandability and comprehensibility of a text. In readability research, only the latter meaning of the word is dealt with [...] Text difficulty can be seen as a synonym to readability. From the reader's perspective, reading proficiency is the corresponding concept." (Ott, 2009, p. 13)

It is possible to define broadly what readability means, since most of the time it is quite obvious to a human reader if a text is generally easy to read. Nevertheless, it is much harder to point at particular factors that lead to this decision. Moreover, recent research has outlined the interaction that occurs when a text is being read, thus introducing external factors that may come into focus.

> "Texts can be difficult or easy, depending on factors inherent in the text, on the relationship between the text and the knowledge and abilities of the reader, and on the activities in which the reader is engaged." (Snow, 2002, p. 14)

Research communities and methods Various communities are concerned with this concept, including language acquisition, text or discourse linguistics, psycholinguistics, web accessibility, language typology, and even research on mathematical entropy, for instance regarding compression algorithms. They correspond to different target audiences, such as children, students, patients, recruits, language learners, and adult native-speakers (sometimes the alleged norm).

At first, readability was seen as a quantification task by the research community. Indeed, the readability formulas provided a numerical estimate, sometimes called a *Readability Index*. which answered a question of the type "How difficult is the text ?". Then, certain thresholds were used to determine if a text is suited for a given grade. Progressively, the thresholds have gained importance, and readability assessment finally has become a classification task similar to language proficiency.

By gathering and assessing the output of a wide range of variables, classification studies group texts according to their difficulty level. This topic is well-studied (Feng, Jansche, Huenerfauth, & Elhadad, 2010; Dell'Orletta, Montemagni, & Venturi, 2011; François & Fairon, 2012) but keeps on being challenging for several reasons.

Salient challenges First of all, a judgment on readability may vary depending on social factors such as education and environment (Snow, 2002), leading to low inter-annotator agreement ratios. The notion of comprehensibility is even more theoretically ill-defined than the notion of readability, which may explain why recent studies focus on empirical results such as benchmarks and evaluation methods.

Secondly, there is matter for discussion regarding the local and global factors that have an impact on readability. In fact, it implies phenomena below and at sentence level, e.g. respectively vocabulary difficulty and syntactic complexity, as well as phenomena occurring over

the whole text, e.g. discourse-level features such as text cohesion and coherence, and anything in between, e.g. the introduction of new words/concepts as well as style-based features (Dell'Orletta et al., 2011). One claim is for instance to go deeper into the simulation of the way one reads, for example by integrating discourse-level features.

The issue of the right level for significant features follows from the lack of distinctive features and acknowledged scientific evidence in linguistics about the way texts are understood. This problem is often addressed by developing a theoretical framework where the empirical results are considered as a mere appliance or validation. In fact, Valette (2008) sees several possible consequences for the association of software tools and theoretical tools: the validation, be it logical or practical, of a model or an instrumentation which deploys the object of research, i.e. what he calls a "virtuous circle".

Thirdly, one may ask how the relevance of a given indicator is assessed. Like all the methodological issues, this evaluation may be highly dependent on the research community and its standards.

Last, the results of a study are not merely an output, they can be a flexible material prone to interpretation distortion and errors. They are also a valuable resource that can be made available to others, raising the question of standards and annotation levels. The existence of well-known annotation schemes such as XML-based text metadata is determining, as it allows for a plausible inspection of scientific results. For a discussion of the open science paradigm, see p. 32.

Steps to solve a classification problem François and Fairon (2012) see three major steps to be performed in order to solve a classification problem, which was for them the design of an AI readability formula:

- Gather a gold-standard corpus,
- Define a set of predictors
- Find the best subset combined with a machine learning algorithm to build a model.

In agreement with this methodology, I would add a proper visualization, interpretation and transmission of the results at the end of it.

2.3.3 State of the art of different widespread approaches

In order to get a glimpse of different facets of this notion as well as of different research methodologies, the five following approaches are presented: a first theoretical one, a linguists' point of view on complexity, a second theoretical approach, psycholinguistics and research on models, the "old school" applied approach to readability, word lists, word frequency and contextual diversity, the recent applied industrial approach, with examples from Amazon, Google, and Microsoft, and finally the case of NLP, which is at a crossing of theoretical and applied approaches.

2.3.3.1 Theoretical approach 1: a linguists' point of view on complexity

Halliday (1992) provides an interesting reference point as the author gives a few insights on the questions one could ask of a given text to find a language model.

One of the points has to do with "text dynamics". Here is how Halliday defines it:

> "It is a form of dynamic in which there is (or seems to be) an increase in complexity over time: namely, the tendency for complexity to increase in the course of the text." (Halliday, 1992, p. 69)

In fact, Halliday develops a very interesting idea from the textual dimension of complexity, also named the "unfolding of the text" (p. 69), its "individuation" or the "logogenesis".[5]

Next he speaks of "some local increases in complexity", for instance in scientific and technical discourse:

> "An example of a more general increase in complexity is provided by grammatical metaphor."
> "*Nominal strings* like glass fracture growth rate and increasing lung cancer death rate tend to accumulate throughout the text rather than being announced at the beginning." (Halliday, 1992, p. 70)

The examples above stand for what he calls a "word complex", which is constructed in the course of the text.

Last but not least, Halliday mentions several measurable criteria: general ones, more specific ones and other possible ways to increase complexity. The first could be captured by lexical density, the number of lexical items per ranking clause, "perhaps weighted for their overall frequency bands"; the second by "the length of nominal chains or adjectival-nominal chains, a selection of verbs which are typically associated with it"; and the third by "the average number of ranking clauses per sentence, the number of phrases and clauses 'rankshifted' inside nominal groups" (Halliday, 1992, p. 71).

There are possibly decreases in the course of a text, but according to him it is harder to see how they could be motivated.

Thus, it can be said that Halliday (1992)'s approach starts from observation and yields a series of potential indicators, not all of which are easily quantifiable, since their main purpose is to contribute to a general model for manually scrutinizing texts.

2.3.3.2 Theoretical approach 2: psycholinguistics and research on models

Psycholinguists may be seen as a discipline which is comparable to linguistics as it has both theoretical and empirical roots. However, the weight of these scientific traditions evolved differently.

Chater and Christiansen (2008) distinguish three main traditions in psycholinguistic language modeling: first, a symbolic (i.e. Chomskyan) tradition; second, connectionnist psycholinguistics; and third, probabilistic models.

The authors state that the Chomskyan approach, as well as nativist theories of language in general, outweighed until recently by far any other one, setting the ground for cognitive science:

> "Chomsky's arguments concerning the formal and computational properties of human language were one of the strongest and most influential lines of argument behind the development of the field of cognitive science, in opposition to behaviorism." (Chater & Christiansen, 2008, p. 477)

[5] The notion of "logogenesis" has apparently been coined by Jay Lemke in *Technical discourse and technocratic ideology*, 1990.

The Symbolic Tradition Chater and Christiansen (2008, p. 479) describe the derivational theory of complexity, i.e. the hypothesis that number and complexity of transformations correlate with processing time and difficulty, as proving "a poor computational model when compared with empirical data". Further work on generative grammar considered the relationship between linguistic theory and processing as indirect, this is how they explain that this Chomskyan tradition progressively disengaged from work on computational modeling. Nonetheless, work on cognitive models has continued, they classify the work of Matthew W. Crocker and Edward Gibson as being in this category.

One of the main issues would be how to deal with the huge local ambiguity of human language.

Connectionist Psycholinguistics In this (more recent) approach, neurons are viewed as entities mapping and transmitting real-valued inputs to each other. The model corresponding to this hypothesis is a connectionist net. According to Chater and Christiansen (2008, p. 481), "'soft' regularities in language [are] more naturally captured by connectionist rather than rule-based method".

The framework of optimality theory is said to take inspiration from both traditions.

Probabilistic Models Still according to Chater and Christiansen (2008, p. 483), "memory or instance-based views are currently widely used across many fields of cognitive science". The fundamental divide in neural network architectures, i.e. between connectionist and probabilistic models, is whether the input is to be processed unidirectionally only or also using a top-down feedback. As far as I know, there are indeed recent trends to improve probabilistic models of reading comprehension, where the experience of the reader is seen as an exposure to statistical regularities embracing phenomena at word or multi-word level, such as collocations, but also sentence features, such as subject-object inversion.

Application to sentence processing Not only the authors agree on the fact that sentence processing has often been thought of in symbolic terms. The shift toward statistical approaches also led to theoretical changes in the way sentence structure is thought of. In fact, according to the authors there have recently been attempts to capture the statistical regularities between words: "'Lexicalized grammars', which carry information about what material co-occurs with specific words, substantially improve computational parsing performance " (Chater & Christiansen, 2008, p. 489). The authors stand for a hybrid approach:

> "There is a variety of overlapping ways in which rule-based and probabilistic factors may interact. [...] The project of building deeper models of human language processing and acquisition involves paying attention to both rules and to graded/probabilistic structure in language." (Chater & Christiansen, 2008, p. 498).

Conclusion The history of psycholinguistic research has a lot in common with the general evolution in linguistics. Probabilistic models coming from cognitive science are comparable to the ones that come from computer science in natural language processing and linguistics. In all cases, there is now a broad consensus toward mixed-model approaches, where comprehensibility for instance is seen both from a grammatical and from a probabilistic perspective.

2.3.3.3 Oldschool applied approach: Word lists, word frequency and contextual diversity

How to build an efficient word list? What are the limits of word frequency measures? These issues are relevant to readability.

First, a word about the context: word lists are used to find difficulties and to try to improve the teaching material, whereas word frequency is used in psycholinguistics as a predictor for cognitive processing load. Thus, this topic deals with education science, psycholinguistics and corpus linguistics.

Coxhead's Academic Word List The academic word list by Coxhead (2000) is a good example of this approach. He finds that students are not generally familiar with academic vocabulary, giving following examples: "substitute", "underlie", "establish" and "inherent" (Coxhead, 2000, p. 214). According to him, these kind of words are "supportive" but not "central" (these two adjectives could be good examples as well).

He starts from principles of corpus linguistics and states that "a register such as academic texts encompasses a variety of subregisters", so that one has to balance the corpus. Coxhead's methodology is interesting. As one can see he probably knows about research by Biber (1993) or Sinclair (1996):

> "To establish whether the AWL [Academic Word List] maintains high coverage over academic texts other than those in the Academic Corpus, I compiled a second corpus of academic texts in English, using the same criteria and sources to select texts and dividing them into the same four disciplines. [...]
> To establish that the AWL is truly an academic word list rather than a general-service word list, I developed a collection of 3,763,733 running words of fiction texts." (Coxhead, 2000, p. 224)

The first test determines if the list is relevant enough, whereas the second one tells if the list is selective enough. Both aim at detecting if it does what it is supposed to do. It seems to be an acknowledged research practice which is still currently used (Grabe & Stoller, 2013).

Word Frequency vs. Contextual Diversity The next research topic I would like to tackle concerns word frequency and word frequency lists. Adelman, Brown, and Quesada (2006) give a good picture of it:

> "It appears that repeated experience with or exposure to a particular word makes it more readable or identifiable. A key assumption of theoretical explanations of the word frequency (WF) effect is that the effect is due to the number of experiences with a word; each (and every) exposure has a long-term influence on accessibility." (Adelman et al., 2006, p. 3)

They distinguish the connectionist models (learning upon each experience of a word) from the lexicon-based ones, where the accessibility of individual lexical entries is governed by frequency. They also refers to the research on memory, in which scholars consider a separation of the exposures in time and context.

They investigate the function of a "contextual diversity", which they define as follows:

> "A normative measure of a word's CD [contextual diversity] may be obtained by counting the number of passages (documents) in a corpus that contain that word." (Adelman et al., 2006, p. 4)

In fact, contextual diversity seems to be a better indicator of reaction times, and thus it may also be relevant for assessing text readability. Their study comes to the following conclusion, where CD stands for contextual diversity and WF for word frequency:

> "In both word naming and lexical decision contextual diversity was more predictive of reaction times than word frequency. Moreover, CD had a unique effect such that high CD led to fast responses, whilst WF had no unique effect or a suppressor effect with high WF leading to slow responses. This implies there is a CD effect, but no facilitatory effect of WF per se." (Adelman et al., 2006, p. 11)

Finally, they infer from their results that they "motivate a theory of reading based on principles from memory research" (p. 13). Adelman et al. (2006) are not the first researchers who study the impact of contextual diversity, but they give a good account of the importance of this phenomenon.

Towards a more efficient word frequency measure Starting from these results, Brysbaert and New (2009) try to provide a more efficient word frequency measure.

Among other interests, they discuss the question of corpus size and type: How big should it be and what kind of texts should be included? In their opinion, "for most practical purposes, a corpus of 16–30 million words suffices for reliable word frequency norms." (Brysbaert & New, 2009, p. 980)

Previous research from them showed that film and television subtitles, as an alternative source of language use, outperformed measures derived from books and Internet searches. What makes subtitles so particular is their vocabulary. They mostly include tangible words (and few conceptional ones). Moreover, long words tend to be avoided.

The next question to arise is whether one should use only lemmas/lemmata or all irregular forms to build the list:

> "Our analyses with the entire Elexicon suggest that, for most practical purposes, lemma frequencies in English are not more informative than WF frequencies. This also seems to be the conclusion reached by Baayen in his most recent articles" (Brysbaert & New, 2009, p. 984)

The authors list the practical implications of the superiority of the contextual diversity measure (Brysbaert & New, 2009, p. 987). Indeed, corpora collected for this purpose have to account for the superiority of the so-called contextual diversity frequency measure.

In their opinion, a corpus consisting of a large number of small excerpts is better than a corpus consisting of a small number of large excerpts. They state that at least 3,000 different samples are needed, with presumably not much gain to be expected above 10,000 samples. Samples of moderate size are required, i.e. a few hundred words to a few thousand words. Additionally, it may not be good to use samples that succeed each other rapidly in time.

As a conclusion, the authors give an idea of the registers to use:

> "The two most interesting language registers currently available are Internet discussion groups and subtitles. [...] On the basis of the English findings, frequencies based on discussion groups seem to be indicated for words longer than seven letters, whereas for short words subtitle frequencies are better." (Brysbaert & New, 2009, p. 988)

In conclusion, the approach of Brysbaert and New (2009) is interesting in the way it revives the sampling issue, which has been highly debated in corpus linguistics (see chapter 1), with empirical arguments. Their frequency models can accommodate text fragments as long as they are not too short or within a too narrow time frame.

I have tested the assumptions of Brysbaert and New (2009) using a German subtitles corpus available for German which has been specially gathered for this task (see chapter 4).

2.3.3.4 Applied industrial approach: Amazon, Google, and Microsoft

Approach and indicators used by Amazon.com to classify books Readability formulas such as the Fog Index, the Flesch Index, and Flesch-Kincaid Index, are apparently used by Amazon: they are mentioned and explained in their text readability help.[6] The formulas are centered on word length and sentence length, which is convenient but by far not always appropriate.

There is another metric named "word complexity", which Amazon defines as follows: "A word is considered 'complex' if it has three or more syllables".[7] One may wonder what happens in the case of proper nouns, for example "Schwarzenegger". There are cases where the syllable recognition is not that easy for an algorithm that was programmed and tested to perform well on English words. The frequency of a proper noun is also interesting per se, because one can expect well-known personalities to be identified much more quickly by the reader. For that matter, named entity recognition may be a key to readability assessment. Besides, not everyone is famous to every kind of reader, yet another reason to use reader profiles.

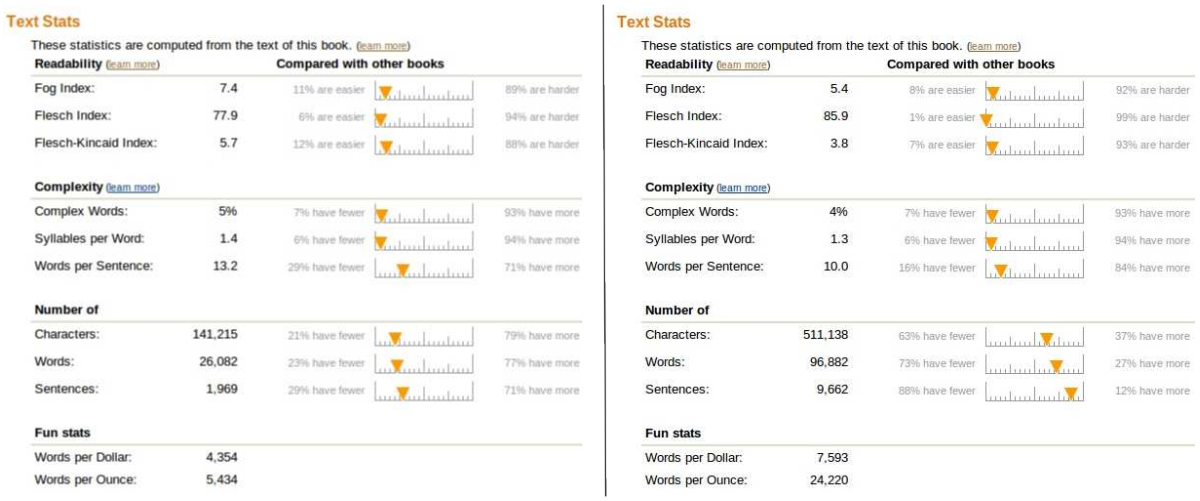
The comparison function is interesting, but only if the categories fit your needs. The example Amazon gives is only half convincing, since the "Children's Books > Ages 4-8" category could deal with books whose content varies a lot. I found no explanation on how the categories were made and how relevant they might be, apart from the fact that the complexity of a text is of an evolving rather than a fixed variable. A book may be easy at the beginning and become more and more complex as it progresses.

Example of Amazon's metrics An example showing particularly well why one cannot rely on these statistics when it comes to get a precise picture of a text's readability.

Figure 2.2 shows two screenshots exemplifying text statistics applied on two very different books, which the metrics fail to describe as such. The two books look quite similar, except for the length of the second one, which seems to contain significantly more words and sentences.

However, the first book is *Pippi Longstocking*, by Astrid Lindgren, a popular classic of childrens literature, whereas the second is *The Sound and The Fury*, a novel by William Faulkner. Neither the genre, nor the time they were published (the first in 1945, the latter in 1929) account

[6] `http://www.amazon.com/gp/search-inside/text-readability-help.html`, as available on 2014-09-01.
[7] Text readability help, *ibid.*



Text Stats

These statistics are computed from the text of this book. (learn more)

Readability (learn more)		Compared with other books	
Fog Index:	7.4	11% are easier	89% are harder
Flesch Index:	77.9	6% are easier	94% are harder
Flesch-Kincaid Index:	5.7	12% are easier	88% are harder
Complexity (learn more)			
Complex Words:	5%	7% have fewer	93% have more
Syllables per Word:	1.4	6% have fewer	94% have more
Words per Sentence:	13.2	29% have fewer	71% have more
Number of			
Characters:	141,215	21% have fewer	79% have more
Words:	26,082	23% have fewer	77% have more
Sentences:	1,969	29% have fewer	71% have more
Fun stats			
Words per Dollar:	4,354		
Words per Ounce:	5,434		

Text Stats

These statistics are computed from the text of this book. (learn more)

Readability (learn more)		Compared with other books	
Fog Index:	5.4	8% are easier	92% are harder
Flesch Index:	85.9	1% are easier	99% are harder
Flesch-Kincaid Index:	3.8	7% are easier	93% are harder
Complexity (learn more)			
Complex Words:	4%	7% have fewer	93% have more
Syllables per Word:	1.3	6% have fewer	94% have more
Words per Sentence:	10.0	16% have fewer	84% have more
Number of			
Characters:	511,138	63% have fewer	37% have more
Words:	96,882	73% have fewer	27% have more
Sentences:	9,662	88% have fewer	12% have more
Fun stats			
Words per Dollar:	7,593		
Words per Ounce:	24,220		

Figure 2.2: Metrics applied to two different fiction books in English, as found on Amazon's website in 2012. To the left *Pippi Longstocking*, by Astrid Lindgren, to the right *The Sound and The Fury*, by William Faulkner. Text length left aside, the metrics fail to reveal any fundamental difference between the two books.

for their different nature. Among other things, Faulkner's novel is considered to be a difficult, skilled experiment with the possibilities and the variety of English language, with a succession of poetic, rich descriptions and transcriptions of Southern dialect, as the action takes place in the state of Mississippi, USA.

Thus, the writing style could not be more different. However, the text statistics make them appear quite close to each other. They also don't acknowledge the diversity of language registers which collide in *The Sound and The Fury*. The criteria used by Amazon are too simplistic, even if they usually perform acceptably well on all kind of texts. The readability formulas that output the first series of results only take the length of words and sentences into account and their scale is designed for the US school system. In fact, the "readability" and "complexity" factors are the same, so these sections are redundant. Nevertheless, it is an interesting approach to try and discriminate between them.

It is clear that the formulas lack depth and adaptability. We need to get a much more complete view of the processes that touch on readability issues.

Still, there may be other reasons that make the books comparable on this basic visualization. At the beginning of *The Sound and The Fury*, the characters are mostly speaking to a child. The ambiguity regarding the sentences and the narrative flow does not make its content that easy to understand, let alone the fact that the described social and interpersonal reality is crude, if not brutal. On the whole, Faulkner's sentences are not particularly short, there are even a few luminous counterexamples, so this may be a failure in the text analysis, for instance in the tokenization process.

Conclusion on Amazon The Amazon text stats give a very general idea of text difficulty. The formulas are robust, they will work on all kind of texts. They are efficient, since the indexes have been developed with a particular eye on teaching levels and children. But these strengths come with two main disadvantages. On the one hand, the generalization of the process is also

its main weakness. It yields approximative results concerning various types of text, which is convenient but sometimes far from being precise. On the other hand, the indexes are based on a scale that fits the US school system, but won't satisfy everyone. The profile issue is not addressed. One cannot tell whether the book contains sentence types or vocabulary one does not like or that one does not understand well.

Filtering search results by reading level: Google The most interesting bits of information which can be extracted from Google's official help page on Google reading levels consist in a brief explanation by a product manager at Google who created the following topic on the help forum: "New Feature: Filter your results by reading level."[8]

Apparently, it was designed as an "annotation" based on a statistical model developed using real word data (i.e. pages that were "manually" classified by teachers). The engine works by performing a word comparison, using the model as well as articles found by Google Scholar.

In the original text:

> "The feature is based primarily on statistical models we built with the help of teachers. We paid teachers to classify pages for different reading levels, and then took their classifications to build a statistical model. With this model, we can compare the words on any web page with the words in the model to classify reading levels. We also use data from Google Scholar, since most of the articles in Scholar are advanced."[9]

It seems to be a model of reading complexity merely based on words. It does not include readability formulas. By comparing the texts to assess with a (gold) standard it aims at being robust.

This model assumes that one doesn't tackle a simple issue using uncommon or difficult words, and that the words are a sufficient criterion. This can lead to curious deformations.

The Googlers think that scientific articles are far out of the linguistic norm. However, the purpose of the authors most of the time is to be as clear as possible, apart from technical words. The identification of such words can be difficult to balance.

Language varieties seem to be taken into account and classified accordingly. For example, a search for "yo mama" returns mostly results qualified as basic, the same is true for "in my hood". It's interesting since it would probably be different if these words were unknown to the system.

On the contrary, the Simple English version of Wikipedia seems to be annotated as intermediate or advanced, although it is meant to be simple, and, in my opinion, it succeeds in doing so, on a lexical as well as on a syntactical and on a semantic level.

To conclude, one may argue that the lack of as well as the rapidly changing nature of technical documentation is typical of the industrial approach. First, in order to keep a competitive edge, not all information is disclosed. Second, the development of applications follows a fast pace, quick and unforeseen readjustments may happen, which adds to the opacity of the system. In that particular case, one may wonder to what extent the model really proves efficient in terms of precision, at a particular moment as well as over time.

[8] It does not seem to have ever been a hot topic, the page does not exist anymore, it has been merged with the general help page about result filtering: `https://support.google.com/websearch/answer/142143`, as available on 2014-09-01.

[9] *ibid.*

Microsoft and social network analysis At the beginning of 2012, Microsoft was planning to analyze several social networks in order to know more about users, so that the search engine can deliver more appropriate results.[10]

Among the variables considered, the “sophistication and education level" of the posts is mentioned. This is highly interesting, because it assumes a double readability assessment, on the reader’s side, and on the side of the search engine. More precisely, this could refer to a classification task.

The extract from the patent quoted below describes how the social network analysis is supposed to work. The text genre in itself is interesting, since it is a highly technical patent text. The numbers refer to sections of a diagram which are not available online.

> “[0117] In addition to skewing the search results to the user’s inferred interests, the user-following engine 112 may further tailor the search results to a user’s comprehension level. For example, an intelligent processing module 156 may be directed to discerning the sophistication and education level of the posts of a user 102. Based on that inference, the customization engine may vary the sophistication level of the customized search result 510. The user-following engine 112 is able to make determinations about comprehension level several ways, including from a user’s posts and from a user’s stored profile. In one example, the user-following engine 112 may discern whether a user is a younger student or an adult professional. In such an example, the user-following engine may tailor the results so that the professional receives results reflecting a higher comprehension level than the results for the student. Any of a wide variety of differentiations may be made. In a further example, the user-following engine may discern a particular specialty of the user, e.g., the user is a marine biologist or an avid cyclist. In such embodiments, a query from a user related to his or her particular area of specialty may return a more sophisticated set of results than the same query from a user not in that area of specialty."[11]

The main drawback of this approach is the determination of a profile based on communication. First of all, people do not necessarily wish to read texts that are as easy (or difficult) as those they write. Secondly, people progress in speaking a language by reading words or expressions they do not already know, so that by doing so Microsoft could prevent young students from developing language skills. Last, communication is an adaptive process: a whole series of adaptations depends on the persons or the group one speaks to, and the sophistication level varies accordingly, which is not necessarily correlated with an education level.

A general example would be that people usually try to become (or to seem) popular on Facebook by mimicking communication strategies other than their own, which involves using shorter sentences and colloquial terms. Another example would be the lack of time, and as a result shorter sentences and messages.

It seems that this strategy is based on the false assumption that one can judge the user’s linguistic abilities by starting from a result that is in fact a construct. In other words, it seems

[10] http://www.geekwire.com/2012/microsoft-idea-deduce-users-mood-smarts-facebook-posts-adjust-search-results/, as seen on 2014-09-01.

[11] http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220120095976%22.PGNR.&OS=DN/20120095976&RS=DN/20120095976, as seen on 2014-09-01.
N.B.: this obviously obfuscated URL makes for a poor content source in the case of web corpus construction.

to be an excessive valuation of performance over competence. There are many reasons why people may speak or write differently in different situations, that is what many sub-disciplines of linguistics are about, and that is what Microsoft is blatantly ignoring in this project.

A reasonable explanation would be that the so-called levels are rough estimates and that the profiles are not fine-grained, i.e. that there are only a few of them. Another explanation may lie in the fact that industrial applications have to become efficient, if not profitable, rapidly, so that not all factors concerned can be taken into account.

2.3.3.5 At the crossing of theoretical and applied approaches: NLP

Tendencies In a recent article about a readability checker prototype for Italian, Dell'Orletta et al. (2011) provide a good overview of current research on readability. The authors offer an extensive review of criteria used by other researchers.

First of all, there is a growing tendency towards statistical language models. In fact, language models are used by François (2009) for example, who considers they are a more efficient replacement for the vocabulary lists used in readability formulas.

Secondly, readability assessment at a lexical or syntactic level has been explored, but factors at a higher level still need to be taken into account. It has been acknowledged since the 80s that the structure of texts and the development of discourse play a major role in making a text more complex. Still, it is harder to focus on discourse features than on syntactic ones.

> "Over the last ten years, work on readability deployed sophisticated NLP techniques, such as syntactic parsing and statistical language modeling, to capture more complex linguistic features and used statistical machine learning to build readability assessment tools. [...] Yet, besides lexical and syntactic complexity features there are other important factors, such as the structure of the text, the definition of discourse topic, discourse cohesion and coherence and so on, playing a central role in determining the reading difficulty of a text." (Dell'Orletta et al., 2011, p. 74)

As a matter of fact, the prototype named READ-IT introduced by Dell'Orletta et al. (2011) does not deal with discourse features.

Features The corpus on which the studies were performed is a sample from the *Weekly Reader*. The OpenNLP chain was used to extract named entities and resolve co-references. Finally, the Weka learning toolkit was used for feature selection and machine learning.

The features of the final selection include:

- Four subsets of discourse features:
  - entity-density features
  - lexical-chain features (chains rely on semantic relations as they are automatically detected)
  - co-reference inference features (a research novelty)
  - entity grid features (transition patterns according to the grammatical roles of the words)

- Language Modeling Features, i.e. train language models
- Parsed Syntactic Features, such as parse tree height
- POS-based Features
- Shallow Features, i.e. traditional readability metrics
- Other features, mainly “perplexity features" according to Schwarm and Ostendorf (2005)

Results According to Feng et al. (2010), combining discourse features doesn't significantly improve accuracy, they do not seem to be useful. Language models trained with information gain outperform those trained with POS labels (apart from words and/or tags alone). Verb phrases appear to be more closely correlated with text complexity than other types of phrases. Noun-based features generate the highest classification accuracy. Average sentence length has dominating predictive power over all other shallow features. The criteria regarding clauses did not perform well, the authors are going to continue working on this.

It is no wonder that those criteria that are simple to implement also perform well. Since they are easier to capture they yield potentially clearer results. On the other hand, it is hard to believe that the discourse features are of no use. More fine-grained features such as these need models that are more accurate, which ultimately means complex models.

> “In general, our selected POS features appear to be more correlated to text complexity than syntactic features, shallow features and most discourse features." (Feng et al., 2010)

The POS-based features are not very detailed, so that I will rather speak of POS-basic features. The authors did not focus on these, although the simple approach apparently found relevant information.

> “A judicious combination of features examined here results in a significant improvement over the state of the art." (Feng et al., 2010)

That leads to another problem: how to be balance the combination? In this study it seems that all the features were equal, but in fact there are always privileged metrics as for instance more and more discourse or word criteria are taken into account. All in all, this particular measurement is prone to be dominated by syntactic indicators, and thus to reflect mostly what happens on the sentential level.

Combination of factors and adaptation In fact, the way indicators are combined is not trivial, and Dell'Orletta et al. (2011) underline the importance of combination:

> “The last few years have been characterized by approaches based on the combination of features ranging over different linguistic levels, namely lexical, syntactic and discourse." (Dell'Orletta et al., 2011, p. 75)

To be able to combine also implies adaptability, which is a key concept, as one has to bear in mind that “reading ease does not follow from intrinsic text properties alone, but it is also affected by the expected audience" (*ibid.*, p. 75).

The authors quote Pitler and Nenkova (2008) as an example of this approach. They also refer to their conclusions on the adaptability of criteria: “When readability is targeted towards adult competent language users a more prominent role is played by discourse features." (*ibid.*, p. 75.)

2.3.4 Common denominator of current methods

2.3.4.1 Finding or defining a gold standard

Inter-annotator agreement As the inter-annotator agreement ratios do not seem to be high enough to be reliable, according to Tanaka-Ishii, Tezuka, and Terada (2010), the number of categories to use for classification is a major issue:

> “Humans are generally unable to precisely judge the level of a given text among 12 arbitrary levels." (Tanaka-Ishii et al., 2010, p. 204)

In order to address this first issue, it is necessary to define precise guidelines or to use existing standards, such as reading competency expected at various school levels or at various stages of foreign language learning. In the first case, expectancies are found to differ between languages and even between countries.

In the second one, the Common European Framework of Reference for Languages is loose enough to offer a general solution, but defining criteria based on its scale may be difficult. In fact, “the annotation of CEF levels of text is not discussed in the key publication" (Ott, 2009, p. 70).

Training data The main drawback of this approach is that the assignment often has to be made by hand. (François & Fairon, 2012) use classifiers trained on texts extracted from school-books and textbooks that are designed for students of French as a Foreign Language and that use the CEFRL grades of difficulty. One has to hypothesize that learning languages at school follows a continuum and that the school books, for example, are coherent (Daoust, Laroche, & Ouellet, 1996). This precisely is the problem that arises when it comes to taking advantage of recent advances in classification, where supervised training methods need fully ordered training data.

Scale of measurement The scale of measurement is an issue by itself, as “the ranges of reading difficulty corresponding to these [school] grades are not necessarily evenly spaced" (Heilman, Collins-Thompson, & Eskenazi, 2008, p. 74). Moreover, between nominal, ordinal, and interval scales, “it is not clear to which scale reading difficulty corresponds" (*ibid.*). Not all classification algorithms can handle the ordinal scale for instance, so that the assumption as to the possibly appropriate scale implies a decision regarding the classifier.

2.3.4.2 Finding possible criteria

Are “deeper" criteria better criteria? Since the end of the 1970s there has been a call for “deeper" criteria:

> “A complete model would include features representing cohesion, informational packaging (e.g. given vs. new information, topic, theme), and rhetorical organization." (Biber, 1992, p. 141)

This situation did not change much, although a lot of progress has been made:

> “Despite the seeming diversity of readability formulas, these measures are all based on, or highly correlated with, two variables: the frequency or familiarity of the words, and the length of the sentences." (McNamara, Louwerse, McCarthy, & Graesser, 2010, p. 3)

There are different ways to look for possible indicators, the main divide probably deals with local and global views. On the one hand, understandability is perceived as a dynamic process, something that evolves in the course of the text. This assumption requires focusing the tools locally (maybe at different scales) on a series of phenomena. On the other hand, the phenomena are considered on a global level, where “global" means in the whole text at once, for instance by calculating means, and not in the whole language system.

Blache (2011) speaks of “difficulty" to clarify that the performance level and not the competence level is meant, as in language complexity studies. He understands global difficulty as the "interpretation difficulty of a sentence or an utterance" and local difficulty as the processing difficulty of a word or a given construction"[12].

Psycholinguistic and linguistic tradition The psycholinguistic tradition refers to the notions of microstructure and macrostructure which were first explained by (Kintsch & Van Dijk, 1978). The first addresses the structure and relations of the individual propositions whereas the latter more generally addresses phenomena at discourse level. (Préfontaine & Lecavalier, 1996) is an example of an attempt to apply this model to comprehensibility issues. In order to analyze the macrostructural level, text sequences are analyzed, most notably by counting the number of so-called reading operations per sequence.

In the fields of psycholinguistics and linguistics it is generally believed that criteria should be meaningful:

> “Complexity values should be computed over strings of lexical items that represent grammatical units, which, as can reasonably be assumed [...] correspond to units that are relevant from a processing point of view." (Kreyer, 2006, p. 45)

NLP tradition In natural language processing, this divide between global and local difficulty implies a theoretical choice that is not always conscious, as interesting tools that come out of the box may be preferred to the development of ad-hoc software. On the other side, there are also researchers who develop a tool to show the power of a given theoretical model.

Most of the time, the indicators rely on a specific methodology and cannot be transferred easily to another framework. This is for instance the case for Latent Semantic Analysis (LSA), which was primarily developed for information retrieval purposes and which is supposed to simulate human knowledge (McNamara et al., 2010).

[12](Blache, 2011, p. 8)

2.3.4.3 Selecting a set of prominent features

Classical formulas Readability formulas have been successful and are still considered to be relevant. The main objections lie with the fact that they are too simplistic. Indeed, they were supposed to be easy to apply to texts and to allow an automatic calculation of the scores later, which was by then a resource-intensive computation:

> "The more variables measured, the more complicated computations would become, and complicated formulas are not practical for classroom teachers to use." (Glazer, 1974, p. 464)

Strikingly, although a lot of different formulas were established for various purposes and various languages, they are all based on the same few indicators:

> "Despite the seeming diversity of readability formulas, these measures are all based on, or highly correlated with, two variables: the frequency or familiarity of the words, and the length of the sentences." (McNamara et al., 2010, p. 3)

Machine learning techniques Nowadays, statistical machine learning is used to determine the most important indicators. On a reasonably sized corpus, the number of features to analyze is not an issue anymore:

> "Over the last ten years, work on readability deployed sophisticated NLP techniques, such as syntactic parsing and statistical language modeling, to capture more complex linguistic features and used statistical machine learning to build readability assessment tools." (Dell'Orletta et al., 2011, p. 74)

The machine learning operation is described by (Heilman et al., 2008) as a "regression of grade level on a set of textual features" and by (Tanaka-Ishii et al., 2010, p. 204) the other way: "Every method of readability assessment extracts some features from a text and maps the feature space to some readability norm." In fact, the first uses logistic regression and yields a possibly weighted score projected on a linear axis whereas the second uses support vector machines that segment the space in order to perform a classification task.

Balancing criteria Two paramount goals in the selection of the "right" indicators deal with the balance of the set. First, redundancy has to be avoided. Either a choice is made between redundant criteria, i.e. criteria that strongly correlate with each other, or their impact on the result is lowered. The second goal is related to the first because it is also a matter of avoiding that criteria artificially get undeserved importance: it is about limiting the fact that indicators were selected and/or given an important weight because they performed well by chance on the test sample. To this end, I have chosen statistical significance tests for my study on web text classification (see chapter 4).

Could a relatively small number of superficial criteria be the answer? At the end of the day, the bigger sets do not appear to be the most efficient ones. (Feng et al., 2010) claims that according to the logistic regression the authors conducted, average sentence length has "dominating predictive power". What is more, selected POS features appear to be more correlated

to text complexity than syntactic features, shallow features and most discourse features. They also show that language models trained on unrelated corpora do not perform better than the selected POS-features. Their features selection process divides the number of features used for the computation by a factor of 10.

The simpler, the better? Salient criteria seem to be more useful than a full range of indicators: "we showed that maximizing the type of linguistic information might not be the best path to go" (François & Fairon, 2012, p. 474). There is no broad consensus in the research community whether all linguistic compartments are needed to assess text readability. Nonetheless, regardless of the number of indicators, diversity might be a plus.

2.3.5 Lessons to learn and possible application scenarios

Towards an applied operationalization All in all, although there is no broad consensus and in spite of the application-based approach, there are convincing results to make use of in further studies, which can then benefit greatly from the diversity of scientific traditions used to tackle issues linked to understandability. The main bias seems to be the extensive operationalization of this notion on little theoretical ground, which is not a decisive drawback as long as no theoretical feedback is expected and which in turn is an opportunity to provide a fruitful instrumental and experimental apparatus.

Genre-related indicators As these indicators deal with the very fabrics of a text, they could prove useful in order to perform other tasks. As a matter of fact, there is evidence that readability is genre-related and that instruments trained on genres help classify texts according to their readability, as (Kate et al., 2010) claims concerning language models.

Accessibility Accessibility is another key concept that seems to be closely linked to comprehensibility. To be able to qualify information in terms of accessibility often implies to classify texts, for example texts gathered on the web:

> "Understandability in general and readability in particular is also an important issue for accessing information over the web as stated in the Web Content Accessibility Guidelines (WCAG) proposed by the Web Accessibility Initiative of the W3C." (Dell'Orletta et al., 2011, p. 73)

Machine reading Comprehensibility also finds a potential use in machine reading, for instance in the case of the DARPA Machine Reading Program:

> "A readability measure can be used to filter out documents of poor readability so that the machine readers will not extract incorrect information because of ambiguity or lack of clarity in the documents." (Kate et al., 2010, p. 546)

Qualification of web texts Thus, the qualification of texts gathered on the web seems to be a potential application for indicators highlighted by readability studies. For the same reason, it is conceivable to apply successful methodological features to this task. The following principles show what conclusions can be drawn.

Vajjala and Meurers (2013) applies readability prediction models on web texts to see if their indicators generalize well. However, the texts taken into account are either newspaper articles

or top search engine results, which are supposed to be very clean from a technical point of view. The authors do not give any hint as to a possible different nature of web texts. All in all, it is a small incursion in the world of corpora from the web.

I present a comprehensive study later in this document (see chapter 4). It takes advantage of some of the features gathered in readability studies, from character to discourse level, and also uses a comparable feature selection methodology based on statistical significance. My results acknowledge the impact of some features on web text qualification, leading to a more accurate handling of web corpus construction.

2.4 Text visualization, an example of corpus processing for digital humanists

2.4.1 Advantages of visualization techniques and examples of global and local visualizations

Two types of visualizations It is not only a matter of scale: the perspective one chooses is crucial when it comes to visualizing how difficult a text is. Two main options should be taken into consideration. On the one hand, an overview in the form of a summary enabling comparison of a series of phenomena for the whole text. On the other hand, visualization taking the course of the text into account, as well as the possible evolution of parameters.

The first type of visualization is exemplified by Amazon's text stats (see p. 91). To sum up, they are easy to read and provide users with a first glimpse of a book, but their simplicity is also their main problem: the kind of information they deliver is not always reliable.

Sooner or later, one has to deal with multidimensional representations as the number of monitored phenomena keeps increasing. That is where a real reflexion on finding a visualization that is faithful and clear at the same time is needed. I would like to introduce two examples of recent research that I find to be relevant to this issue.

An approach inspired by computer science The first one is taken from an article by Oelke, Spretke, Stoffel, and Keim (2010). It is part of a work by computer scientists trying to apply their approach and their experience with visualization to language processing. The result can be seen on Figure 2.3.

It is quite obvious that the summary is obfuscated by the amount of information. As far as I know, this kind of approach grounds on the belief that the human eye is able to extract patterns from a complex image. This idea seems to date back to Bertin (1967), a seminal contribution to information visualization. But in this case the patterns do not always appear, as the reader cannot not easily control the display of information, which may hinder him at discerning patterns.

Nonetheless, there are two interesting aspects to this work, as the authors start from readability formulas comparable to those used by Amazon (see p. 91). The difference is that they try to mirror possible evolutions within the text in an intuitive way.

A visualization following the course of the text The second example comes from Karmakar and Zhu (2011). It features an attempt to take advantage of text structures to reach a relevant

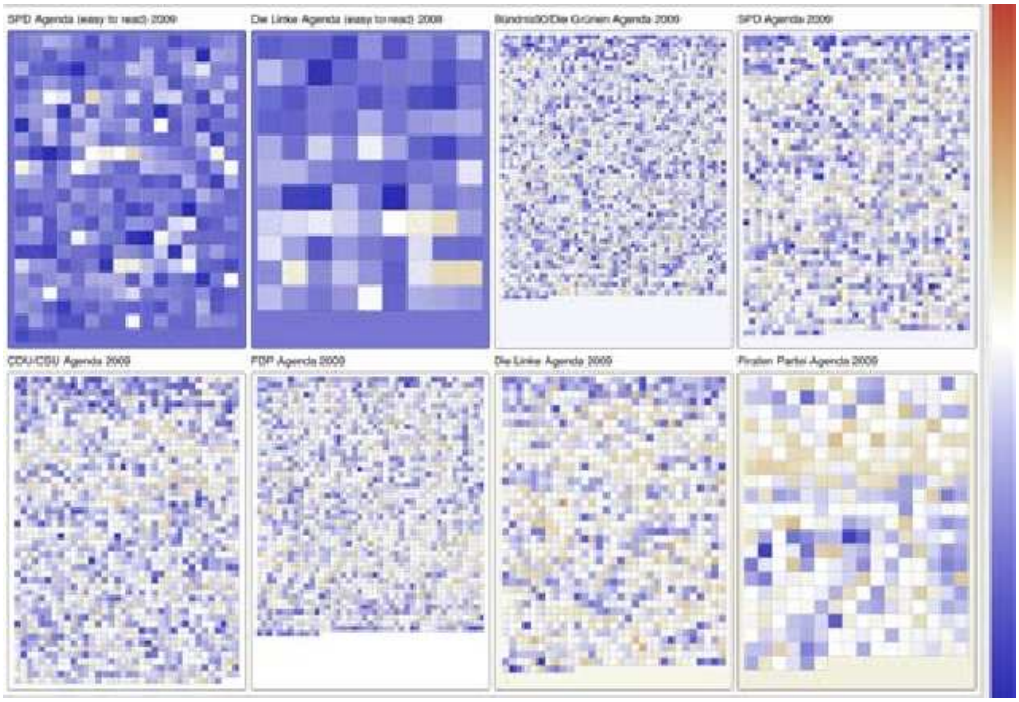


Figure 2.3: A visualization of several German political programs in terms of text complexity. Blue areas signalize text parts which are supposedly easy to read. The texts by the SPD and Die Linke on the left side of the first row are considered easier to read in the study.

image of a text. In this regard, their approach is closer to natural language processing, its operating model, and its evaluation metrics.

It is a local approach of readability as the text is seen in a linear fashion, as Figure 2.4 shows. The horizontal bars on this chart picture the sentences of the article, so that a relevant criterion immediately strikes the eye: the evolution of sentence length in the text becomes clear. So does the distribution of complex words and complex parts (as they use tools to estimate parse tree depth) in the text.

I will not discuss the criteria which Karmakar and Zhu (2011) use to assess readability here, I would just like to point out that their definition of complexity is not necessarily in accordance with linguistic standards and that their interpretation of the visualizations is not obvious. Especially their views on clause complexity are dubious, since such complexity is only rendered imperfectly by figure 2.4, which gives no indication of syntactic features of readability.

Common features There are a few common features between these two research papers which seem to be good starting points.

First of all, in both cases, difficulty is shown by color intensity, where darker (or more intense) signifies a more complex passage.

Second, a visualization of this kind may include several scientific frameworks and in fact heterogeneous indicators. In order to draw relevant conclusions one has to be aware of these differences. They could also be shown, for instance by letting the user select the framework or the indicators that appeal to him.

Third, the ability of the human eye to distinguish between various phenomena and to

Figure 2.4: Linear visualization of the sentences from a newspaper article in English. The lines are sentences, and the boxes are words. The darker the box, the more complex the word.

apprehend color differences (Bertin, 1967) should not be under- nor overestimated. There may be an optimal degree of visual complexity.

2.4.2 Visualization of corpora

2.4.2.1 Introduction

Corpus exploration as a goal One may consider that basic visualization techniques are already used in corpus linguistics, since concordances, collocation networks, or key word clouds are ways to see through a corpus (Rayson & Mariani, 2009). Nonetheless, there is still a lot of work to do to catch up on the computer science field of information visualization, as Rayson and Mariani (2009) acknowledges:

> "We wish to allow linguists to explore their data in 'strange' new ways and to seek out new patterns and new visualizations." (Rayson & Mariani, 2009)

In fact, exploration is a frequently used keyword when it comes to make corpus content available through visualization, which cannot be reduced to a mere statistical analysis but grounds on more complex processes:

> "Statistical tools alone are not sufficient for 'distant reading' analysis: methods to aid in the analysis and exploration of the results of automated text processing are needed, and visualization is one approach that may help." (Collins, Viegas, & Wattenberg, 2009)

Imaginative methods Imagination is another keyword, especially concerning digital humanities and arts, as there is both a need and a real potential concerning graphic aids for digital humanists, with text objects being at a crossing between quantitative methods and aesthetics:

> "A great many of the visualization methods applied to text are derived from analytical quantitative methods that were originally borrowed from the sciences. This is an interesting area of application because there are also other more imaginative visualization tools that owe more to the arts than the sciences." (Jessop, 2008)

By taking recourse to a more creative potential when exploring corpora, the link is made to the role of serendipity in corpus analysis, which has been said to be a major drive of corpus linguistics (see p. 41).

Brief history It has been clear from early on that the notion of web genre can be useful regarding classification and visualization (Bretan, Dewe, Hallberg, Wolkert, & Karlgren, 1998). One of the first attempts to apply visualization techniques to texts was the "shape of Shakespeare" by Rohrer, Ebert, and Sibert (1998). Clustering methods were used to let set emerge among textual data as well as metadata. It may seem rudimentary by today's standards or far from being a sophisticated "view" on literature but the "distant reading" approach is precisely about seeing the texts in another perspective and exploring the corpus interactively. Other examples of text mining approaches enriching visualization techniques include the document atlas of Fortuna, Grobelnik, and Mladenic (2005), and the parallel tag clouds of Collins et al. (2009).

The criticism concerning culturomics seems to hold true for corpus visualization as well: there is still a gap to bridge between information visualization, NLP, and digital humanities. The exploration of digital text collections obtains better results and reaches a larger user base if work on visualization is conducted in a dialog between philologists and NLP experts.

2.4.2.2 Interfaces are necessary, the example of visual analytics

Why use visual analytics in linguistics? Interfaces are necessary to get a glimpse of linguistic data. Tanguy (2012) explains why and on what conditions data visualization could help linguists. He gives a few reasons for using the methods from the emerging field of visual analytics and mentions some of its upholders, like D. Keim in Germany or J.-D. Fekete in France. But he also states that the methods are not well adapted to the prevailing models of scientific evaluation.

His main point is the (fast) growing size and complexity of linguistic data. Visualization comes in handy when selecting, listing, or counting phenomena do not prove useful anymore. There is evidence from the field of cognitive psychology that an approach based on form recognition may lead to an interpretation. Briefly, new needs come forth when calculations come short.

Tanguy (2012) gives two main examples of cases where it is obvious: firstly the analysis of networks, which can be linguistically relevant i.e. for dependency relations within a sentence or a text, and secondly the multiple characteristics conferred to individual data, say the multiple layers of annotation.

He sees three main goals in data analysis that may be reached using visualizations:

- to construct a global point of view (like an aerial view)

- to look for configurations
- to cross data of different natures (one could also say on different scales)

What is still to be done if this method is to be adopted? Nonetheless, the notion of visualization by itself is not a solution to any given problem, one has to find the processes best adapted, which in turn are a construct, a limited projection of the complexity of the data.

It is thus important to leave the users room for experiment and trial and error. A few valuable insights may only appear if visualization parameters are allowed to vary. Tanguy suggests three kinds of evolutions:

- the selection of the dimensions to display and their mode of representation
- a whole series of operations on the constructed view
- last, a fine-tuning of both visualization and data

In this respect, Tanguy quotes Ben Shneiderman's mantra, an influential scholar in the field of information visualization: "Overview first, zoom and filter, then details-on-demand".

The last problem may lie in the complexity of the visualization tools. Tanguy sees three main abilities needed to deal with this matter[13]:

- deep as well as fine-grained knowledge of the analyzed data
- experience with the visualization processes
- competence in data analysis

2.4.2.3 Interpretation, evaluation and visualization of the results

With all the artificial intelligence involved in text classification, decisions are sometimes made on an abstract level that has more to do with algorithm tuning than with linguistics. There are cases where interpretation falls short:

> "In sum, then, we are left with extremely sophisticated metrics that produce extremely precise values which can only be interpreted extremely imprecisely." (Kreyer, 2006, p. 53)

Interpretation as well as evaluation are key steps in order to decide if results confirm the expected norm. Confronted with a lot of features, it may be difficult to get a glimpse of the big picture without a proper visualization of the processes at stake. Indeed, visualization comes in handy when selecting, listing or counting phenomena does not prove useful anymore (Tanguy, 2012). It could be useful to be able to observe the impact of different factors.

[13]There seem to be frequently three subcomponents in Tanguy's demonstrations, an interesting twist typical of French academic culture.

2.5 Conclusions on current research trends

2.5.1 Summary

2.5.1.1 Methodology

In natively digitized text forms, text quality variation throughout the corpus is a major issue, not so much in terms of typos or diverging orthographic phenomena, but rather in terms of content. Indeed, automatically generated text and spam are generally clear cases of unwanted text in a linguistic corpus. These texts at least would need to be marked as such, which necessitates being able to identify them.

I have presented three different perspectives on web document quality analysis: specific web text quality issues, readability assessment, and visualization. Research fields in computer science are well-represented throughout these topics, which is encouraging since it means that there is a real effort towards the improvement of web document collections.

The issue is usually seen as a text classification problem, which involves two main distinct classification methods: on the one hand “knowledge-poor" approaches relying on machine learning algorithms, and on the other hand expert approaches based on knowledge on the side of the researcher. Since machine learning techniques are currently very popular, the classification process boils down to finding the right textual clues, which implies empirical testing and/or expert knowledge.

This knowledge can be found in terms of the selection of cues or, more frequently, in terms of annotation in supervised methods, where training data are used to foster a classification.

2.5.1.2 Guiding principles

Steadiness is worthwhile For the annotation of a gold standard, precise guidelines are valuable even if they are not perfect and do not suit all purposes. Steadiness is necessary in order to get a better idea of the classification task, on the researcher’s side, and better results for e.g. machine learning, on the application’s side.

Stay on the surface One criterion cannot be dismissed just because it seems too simplistic: surface features seem to be efficient. They also enable us to go through the texts at a faster speed as fewer operations are performed. This could be decisive as the number of texts to qualify becomes larger.

The methods used in order to assess the statistical significance of the indicators could easily be adapted.

A global approach to the texts In my opinion, the global vs. local divide which exists in text classification and visualization does not quite apply to web corpus construction, since whole texts are being considered. The qualification of web texts is roughly a classification task where the whole text needs to be “read" before making a decision.

However, it is imaginable to use visualization tools in order to detect local uncertainties. Local characteristics can also be used as disqualifying features such as black lists for spam and text profiles such as list formats.

What annotation and visualization have to offer Annotation/constructed indicators and visualization are both equally important. Using annotation levels or suited indicators, a tool can provide an insight on the nature, structure, and fabric of texts. This can allow for a requalification that would suit other purposes, even if contrary, like an analysis of spam on the internet, or more precisely, like the construction of a subcorpus on a specialized issue.

2.5.2 A gap to bridge between quantitative analysis and (corpus) linguistics

Approximations and abstract indicators The greater part of the qualification and classification work presented in this chapter is of an applied nature, and does not seem to concern linguistics, or even corpus linguistics, directly. The classification operation grounds on an approximation rather than an actual “reading" of the texts: it takes place on an abstract level, where features of numerical nature have already been extracted from the texts. Therefore it is important to bear in mind that what is being classified are series of features rather than texts, and potential improvements may rather stem from more adequate features than from a better classification algorithm.

A positive perspective on quantification For William Thompson, better known as Lord Kelvin, knowledge can be equaled to quantification:

> “I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."
> William Thompson, Lecture on “Electrical Units of Measurement" (3 May 1883)

This perspective on measurement can be found among computer scientists and researchers in NLP. In fact, the principle of encoding data in an abstract fashion so that it can be processed is so prominent that it may even supersede the usual distinction between grammar-based and statistical approaches:

> “Many people now believe there are only two approaches to natural language processing:
>
> - a deep approach that relies on hand-coded grammars and ontologies, represented as complex networks of relations; and
> - a statistical approach that relies on learning n-gram statistics from large corpora.
>
> In reality, three orthogonal problems arise:
>
> - choosing a representation language,
> - encoding a model in that language, and
> - performing inference on the model.

(Halevy et al., 2009, p. 9)

Thus, for Halevy et al. (2009), building a grammar or finding features in texts can be integrated under the same assumptions into a bigger picture, that of formalization. The abstract level is that of the representation language, which raises questions as to how extensive and powerful the resulting models can be.

A skeptical perspective on quantification Nonetheless, other scientific traditions adopt a rather skeptical stance regarding measurement and construction of indicators based on numerical values. Gaston Bachelard, in his first essay, on the topic of "approached knowledge" (Bachelard, 1927), explains that measures cannot be considered for themselves. According to him, the fact that a measure is precise enough gives us the illusion that something exists or just became real. He even states "Do you want to believe in reality? Measure it."[14]

Bachelard criticizes what he calls "the strange equivalence that modern science has established between measurement and knowledge."[15] He militates in favor of an approached knowledge, i.e. a form that is simple enough to embrace a phenomenon without having to be too close to the results obtained by measures. Avoiding unnecessary variables, a formula gains in inner coherence. The difficulty is to get from the fact to the type.

He argues that knowledge often starts by an intuition[16]. In a similar perspective, an interview with children books' author Sabine Ludwig[17] reminded me that readability checking cannot capture the inventive talent. It fails to take into account a sense of writing that makes ideas or stories easier to understand out of seemingly complex or long text parts. I also realized that not all genres are equally measurable. Narration in particular is based on interwoven features for which there is hardly a grid to be designed.

A methodological balance in a technological environment The suspicion concerning data-intensive methods and quantification is roughly similar to the opposition to big data and "big corpora" in part of the corpus linguistics community. However, since the research environment is increasingly a technological environment, as described in chapter 1, there is room for opportunistic approaches.

The methodological insights of other disciplines show that there are decisions during web corpus building which are not made by humans, contrary to the careful selection of texts in traditional reference corpora. More generally, at the time of web for corpus, we live in an environment which is spanned by technological innovation and artificial intelligence. In web corpus construction as elsewhere, machines can be much faster without sacrificing too much accuracy.

NLP is at the crossing of several research traditions, but usually adopts a resolutely applied approach, where efficiency and applications count. The emphasis is put in the optimization of

[14] "Et pourtant, que ce soit dans la mesure ou dans une comparaison qualitative, il n'y a toujours qu'un jugement sur un ordre : un point vient après un autre, une couleur est plus ou moins rouge qu'une autre. Mais, la précision emporte tout, elle donne à la certitude un caractère si solide que la connaissance nous semble vraiment concrète et utile; elle nous donne l'illusion de toucher le réel. Voulez-vous croire au réel, mesurez-le." (Bachelard, 1927, p. 52)

[15] "L'étrange équivalence que la science moderne a établi entre la mesure et la connaissance." *ibid.*, p. 54

[16] *ibid.*, p. 149

[17] `https://de.wikipedia.org/wiki/Sabine_Ludwig`

algorithms rather than in the cleaning of the data set or finding the “right" data. There seems to be a consensus where researchers admit that superficial criteria are surprisingly efficient.

It may be because the measurements undertaken by researchers are about detecting the most salient features, which even from an algorithmic point of view makes things easier. But it may also be a scientific apparatus problem residing in a lack of instruments which are sensitive or fine-grained enough to register more subtle features, for instance on discourse level.

On the other hand, there is in linguistics and psycholinguistics a notion of meaningful, comprehensible and reproducible discrimination, as well as a particular attention, if not literally a suspicion, concerning the instrumental and experimental apparatus.

To conclude, if a balance is to be found, it may not be a compositional but rather a methodological one, between application and theoretical concerns, thus bridging a gap between technological ground and theoretical linguistics. This involves bringing more techniques of NLP to linguistics, but also involves acknowledging that there are issues which require qualitative, fine-grained approaches (Habert, 2000), for instance tasks which are not easily decidable, which may result for instance in a low inter-annotator agreement (see chapter 4 for an example).

Achievements used in the following chapters Some of the methodological insights and achievements described in this chapter are applied in chapter 4, with work on web corpus construction and evaluation of the gathered material.

Since text quality assessment is a real-world problem, features used by proponents of NLP and computer science approaches are used to qualify corpus sources and web texts. More specifically, URL qualification involves avoiding low-quality content to find reliable sources, while text qualification benefits from the use of feature selection techniques used in readability assessment.

Last, visualization of corpus features is used as a way to benchmark the content and compare web corpora with established reference corpora.

Chapter 3

From the start URLs to the accessible corpus document: Web text collection and preprocessing

"Probe the universe in a myriad points."

— Henry D. Thoreau, *The Journal of Henry David Thoreau*, Boston: Houghton Mifflin Co., 1906, p. 457.

3.1 Introduction

3.1.1 Structure of the Web and definition of web crawling

The Web The Web is ubiquitous in today's information society. As the fundamental processes governing its constitution and its functioning prove to be robust and scalable enough, it becomes more and more a "world wide web". These principles are well-known and well-studied, they are not a matter of chance. However, how people use it, what they do with websites in terms of interaction, creation, or habits bears a huge potential in terms of scientific study. This is particularly the case for social sciences, as a part of what Hendler, Shadbolt, Hall, Berners-Lee, and Weitzner (2008) call Web science. The exact structure of the WWW cannot be exactly determined, thereby leaving room for studies in computer and information science.

Since a full-fledged web science study falls beyond the scope of this work, I will focus in the following on technical aspects which make up the very essence of a web crawl and I will describe the challenges to be faced when text is to be gathered on the Web.

First of all, concerning the basic mechanisms, a crawl grounds on the notion of URIs (Uniform Resource Identifiers) and their subclass URLs (Uniform Resource Locators). They are what Berners-Lee et al. (2006) call "basic ingredients" of the Web, making the concept of a web crawl even possible:

> "The Web is a space in which resources are identified by Uniform Resource Identifiers (URIs). There are protocols to support interaction between agents, and formats to represent the information resources. These are the basic ingredients of the Web. On their design depends the utility and efficiency of Web interaction, and that design depends in turn on a number of principles, some of which were part of the original conception, while others had to be learned from experience." (Berners-Lee et al., 2006, p. 8)

The robustness of the URL scheme fosters to a great extent the robustness of the Web itself. As in the expression "surfing the Web", how the Web is perceived is determined by the experience of following links that actually lead from website to website, being able to record them and make the "surf" reproducible because technically secure.

From a more technical point of view, using notions familiar to computer science, the image of a graph with nodes and edges is commonly used to describe the Web:

> "One way to understand the Web, familiar to many in CS [Computer Science], is as a graph whose nodes are Web pages (defined as static HTML documents) and whose edges are the hypertext links among these nodes. This was named the 'Web graph', which also included the first related analysis. The in-degree of the Web graph was shown [...] to follow a power-law distribution." (Hendler et al., 2008, p. 64)

Zipf's law in linguistics is an example of power law, a functional relationship between two quantities where one quantity varies as a power of another. It states that the frequency of any word is inversely proportional to its rank in the frequency table, is considered to be a power law, where frequency of an item or event is inversely proportional to its frequency rank.

The web is thought to be a scale-free network, as opposed to a random network (Schäfer & Bildhauer, 2013, p. 8) because of link distribution. In other words, the global shape of the web graph is probably not random, there are pages who benefit a lot more from interlinking than others, and many cases where the links only go in one direction.

> "Each node has an in-degree (the number of nodes linking to it) and an out-degree (number of nodes linked to by it). It is usually reported that the in-degrees in the web graph are distributed according to a power law." (Biemann et al., 2013, p. 23)

By way of consequence, one may think of the web graph as a polynuclear structure where the nuclei are quite dense and well-interlinked, with a vast, scattered periphery and probably not too many intermediate pages somewhere in-between. This structure has a tremendous impact on certain crawling strategies described below.

The problem is that there are probably different linguistic realities behind link distribution phenomena. While these notions of web science may seem abstract, the centrality and weight of a website could be compared to the difference between the language variant of the public speaker of an organization, and the variants spoken by various members.

Possible ways to analyze these phenomena and to cope with them are described in the experiments below (cf chapter 4).

Web crawler The basic definition of a web crawler given by Olston and Najork (2010) gets to the point:

> "A web crawler (also known as a robot or a spider) is a system for the bulk downloading of web pages." (Olston & Najork, 2010, p. 176)

As such, crawling is no more than a massive download of web pages. However, since the Web is large and diverse, building web crawlers has evolved into a subtle combination of skills.

Web crawling The starting ground that makes web crawling possible in the first place is the connectedness of the web and the existing standards concerning the presence of links in the form of Uniform Resource Locators (URLs) (Schäfer & Bildhauer, 2013, p. 8).

> "The raison d'être for web crawlers lies in the fact that the web is not a centrally managed repository of information, but rather consists of hundreds of millions of independent web content providers, each one providing their own services, and many competing with one another." (Olston & Najork, 2010, p. 176)

Probably because the web has been and continues to be shaped by computer science paradigms related to concepts for theoretical reasons and linked to efficiency for practical reasons, a crawl (web crawling operation) is most commonly seen as the traversal of a web graph.

> "Crawling (sometimes also called spidering, with occasional minimal semantic differences) is the recursive process of discovering and downloading Web pages by following links extracted (or harvested) from pages already known." (Schäfer & Bildhauer, 2013, p. 15)

> “Web crawling is the process of fetching web documents by recursively following hyperlinks. Web documents link to unique addresses (URLs) of other documents, thus forming a directed and cyclic graph with the documents as nodes and the links as edges." (Biemann et al., 2013, p. 23)

One may risk an analogy with space exploration by saying that in both cases the size, composition, and shape of the universe cannot be determined with absolute certainty. It involves a conceptual effort, a theoretical framework, as well as huge datasets of measurements in order to enable scientists to get an indirect idea of the characteristics of their universe.

Let us say for now that there are web pages which will most probably be found with any kind of crawling strategy whatsoever as well as nearly regardless of the length of the crawl, while others may be as “interesting" as the first ones but still won’t be found even by extensive strategies.

3.1.2 “Offline" web corpora: overview of constraints and processing chain

Interest The interest of “offline" web corpora is described in detail in the first chapter. The web document retrieval phase, which leads to “offline", linguistically processed corpora, is seen as a remedy to Googleology (see p. 43), i.e. approximations using tools which are not primarily designed for linguistic research.

First, it enables researchers to access a considerable amount of internet content in their target language and/or sometimes text type. Second, this approach does not have the pitfalls of direct search engines queries regarding linguistic processing, since basic annotation and queries at least are provided with the end product, i.e. the accessible corpus.

Goals The main possible goals of web corpus construction can be summed up as follows:

1. Find content sources
2. Select linguistically relevant web documents
3. Provide rich metadata
4. Remove uninteresting parts (or noise)
5. Republish the content, or make it accessible

These goals are rather general, emphasis may lie on some of them more than others, according to the background of the research project. Additionally, not all goals are practically attainable. Some are even conflicting with each other, there are for example content sources which are especially interesting but do not allow for a republication.

Typical processing pipeline Reaching the goals involves setting up a whole processing pipeline, which goes from sources discovery to an exploitable corpus. In order to provide a basic overview of the whole process evoked in this thesis, important steps could be listed as follows:

1. Find the web documents

2. Download the files
3. Crawl (if applicable)
4. Process the documents
5. Annotate the text
6. Index
7. Run tests on the corpus

This chapter will tackle up-to-date methods to collect and prepare web documents. More precisely, the crawling steps will be described in more detail p. 120, while the processing steps will (also) be further elaborated p. 131.

Skills ideally needed Since web corpus construction is both a complex and diverse enterprise, the skills that are ideally needed in order to conduct it as reliably as possible are manifold:

- Knowledge of Web structure and latest evolutions, most notably in order to find interesting spots and efficient ways through web space;
- Efficient, scalable programming, in order to allow for web-scale computations on hardware that is often not high end by industry standards;
- Text scrutiny, since a special interest and attention for properly prepared texts is needed;
- Knowledge of information architecture, as both web page gathering as well as content publishing involve a clear vision of structured content.

Similarly to the goals, not all the skills are imperatively required, nor can they necessarily be mastered by a single person or even a specific team.

3.2 Data collection: how is web document retrieval done and which challenges arise?

3.2.1 Introduction

3.2.1.1 Challenges

First of all, where does one find, for instance, “German as spoken/written on the web"? Does it even concretely exist or is it rather a continuum? Considering the ongoing shift from web *as* corpus to web *for* corpus, mostly due to an expanding web universe and the potential need for a better text quality, it is obvious that only a small portion of the German web space is to be explored.

There are two main types of data collection processes, mostly corresponding to both web corpus approaches described in the last chapter: restricted retrieval for a special field of interest such as a particular website, and web crawling to allow for web traversal and website discovery.

Web crawling techniques can also be applied to a restricted target in order to look for particular text characteristics or even text types. Thus, if the circumscribed retrieval corresponds to "corpora from the web" in a narrow sense, web crawling does not necessary imply a non-finite or general-purpose approach, as we will see in the following.

Concerning the state of the art, due to the number of industrial interests in information retrieval, there are many existing (very) large document collections gathered from the web. However, most of the time they have not been documented in order not to lose a competitive edge. Even in the case of Google, with probably the best documented crawling processes in terms of research articles, information stays scattered.

This is all the more true concerning web corpora, so that Baroni et al. (2009) declare:

> "We are not aware of much publicly documented work on developing large-scale, general-purpose web-derived corpora." (p. 210)

A thorough, open documentation of the whole crawling process has only been available since 2009 or 2010, for instance with the work of Olston and Najork (2010).

3.2.1.2 Big data opportunism and its caveats

There seems to be a "quick and dirty" work tradition concerning web corpora, coming especially from language modeling and machine translation specialists. N-gram-based approaches are no panacea, since a proper cleaning and segmentation stays crucial:

> "Web1T5[1] shows that sheer size cannot make up for 'messy' content and lack of annotation." (Biemann et al., 2013, p. 43)

Researchers in the machine translation field have made another attempt to outsource competence and computing power, using data gathered by the CommonCrawl project[2] to find parallel corpora (Smith et al., 2013). Purely URL-based approaches to select texts as used in this approach favor speed over precision or sacrifice precision for speed, which masks a series of problems concerning data quality. Indeed, language identification tasks are a good example of this phenomenon (Baykan, Henzinger, & Weber, 2008). Machine-generated text and similar issues are also a problem, even if one argues that training a machine translation system on machine-translated data is not a problem as long as the results improve.

Indeed, the article by Smith et al. (2013) is actually symptomatic for a general trend, since it involved well-known researchers from several institutions. This trend is focused on practical application and benchmarks such as the translation evaluation frameworks. The ability to gather as much parallel text as possible to run through existing toolchains and finally score through evaluation frameworks supersedes a proper quality evaluation phase.

There is a widely-shared belief among computational linguists working on web corpora that researchers could and should take advantage of existing linguistic processing and annotation infrastructure:

> "Unless web pages are carefully selected, cleaned, de-duplicated and enriched with linguistic annotation, even a gigantic text collection such as Web1T5 offers little

[1] Google's Web 1T 5-gram dataset from 2006 https://catalog.ldc.upenn.edu/LDC2006T13
[2] http://commoncrawl.org/

> value for linguistic research. In some tasks, web corpora are still inferior to traditional corpora (and even to other online material such as Wikipedia articles), but this is likely to change with a further increase in corpus size." (Biemann et al., 2013, p. 45)

One may remark that this statement is in a way contradictory, because the authors state that the need for more precise preprocessing tools could be compensated by size in a comparison with traditional (i.e. non-web) corpora regarding corpus quality.

Still, the debate on quality highlights how necessary it is to develop a processing chain adapted to the constitution of web corpora. In a way, the enthusiasm about huge datasets and the belief that probabilistic approaches can mitigate potential design flaws and inconstant text quality can become an obstacle to corpus approaches on the linguistic side. That is why approaches to web corpora by linguists or at least for linguists are needed.

3.2.2 Finding URL sources: Known advantages and difficulties

3.2.2.1 The notion of URL seeds

URLs are called seeds if they are used to start a crawl. The URLs seen during the crawl and added to the list of URLs to visit are called the frontier.

There is a difference between URL sources and URL seeds: while a source may be used as a seed, it is not necessarily possible to crawl a source such as a search engine.

Be it sources or seeds, relevant URLs are crucial for web crawling and ultimately web corpus construction. Obviously, they have a huge impact on a crawl's trajectory, and it is widely known that special attention should be taken in order to allow for a good start. In the case of general-purpose web corpora, "good" mainly means diverse and representative enough:

> "The first step in corpus construction consists in identifying different sets of seed URLs which ensure variety in terms of content and genre." (Baroni et al., 2009, p. 213)

Thus, finding URL seeds should not be considered a trivial task. It is mostly done by selecting the right sources for the right task. The following subsections sketch a few established techniques for this.

3.2.2.2 Specialized sources

First of all, in the case of specialized sources, finding the URLs of documents to download is relatively easy, since the target website is known in advance, and since the target URLs are sometimes even picked manually.

Newspaper websites are a good example of specialized corpora already used by pre-Web researchers. They are known to possess interesting characteristics for linguistic study of a widespread if not standard variant.

For example, Hoffmann (2006) builds what he calls a "specialized corpus of spoken data" made of publicly available CNN transcripts retrieved from the website cnn.com. There is no particular section dedicated to the finding of sources, simply because it is not considered to be an issue.

This is not systematically the case, there are specialized sources for which the gathering of URL sources implies more effort than a mere URL extraction in a homepage, which would be the default, as shown below.

Additionally, the case of URL directories is also described below.

3.2.2.3 General-purpose track: the BootCaT method

Introduction Ten years after the seminal article describing the approach, the BootCaT method (Baroni & Bernardini, 2004) can still be considered as the current method of finding seed URLs. It consists in repeated search engine queries using several word seeds that are randomly combined, first coming from an initial list and later from unigram extraction over the corpus itself.

By this method so-called seed URLs are gathered which are in turn used as a starting point for web crawlers. This approach is not limited to English: it has been successfully used by Baroni et al. (2009) as well as by Kilgarriff et al. (2010) for major world languages.

The main purpose of the BootCaT method is twofold: on the one hand, it enables linguists to build corpora in a convenient fashion, and on the other hand, it makes corpora available "offline", i.e. independently from the websites they were gathered from.

> "Google is used to obtain a list of documents, but then these documents are retrieved and post-processed by the researcher (tokenized, POS-tagged, etc.) locally, so that the stability of the data will no longer depend on Google, the researcher has full access to the corpus and, with the appropriate tools, the corpus can be interrogated with sophisticated linguistic queries." (Baroni & Ueyama, 2006, p. 33)

This approach has been shown to be productive. However, its simplicity masks a number of questions which remain unanswered, and the apparent stability of the end product does not even imply a relative stability of the techniques used to get to the web documents.

Practical shortcomings Until recently, the BootCaT method could be used in free web corpus building approaches. To my best knowledge it is now passé because of increasing limitations on the search engines' APIs, which make the querying process on a low budget much slower or impossible. Other technical difficulties include diverse and partly unknown search biases due in part to search engine optimization tricks on the side of the content publishers, as well as undocumented adjustments of sorting algorithms on the side of search engineering.

All in all, the APIs may be too expensive and/or too unstable to support large-scale corpus building projects. Tanguy (2013) even claims that collaboration with search engines has become strictly impossible due to the commercial context, which impacts all web corpus approaches.[3]

Additionally, API changes are combined with an evolving web document structure. Much more content than before is fetched from other sources at the time the main page is requested. Well-known examples are Twitter or RSS feeds included in a page, as well as various kinds of embedded content such as videos or pictures. While it is not that frequent, there are other kinds of inclusions of "timelines" containing text, for instance newspapers websites which link to the most read articles of a partner website, with a title, a description, or the beginning of the article.

[3] "Les enjeux économiques qui entourent les moteurs de recherche sont désormais tels que la collaboration devient simplement impossible. Les fermetures de ces services mettent à mal l'ensemble des approches [...] de la création d'un corpus à la volée jusqu'à l'interrogation indirecte et enrichie des moteurs." (Tanguy, 2013, p. 14)

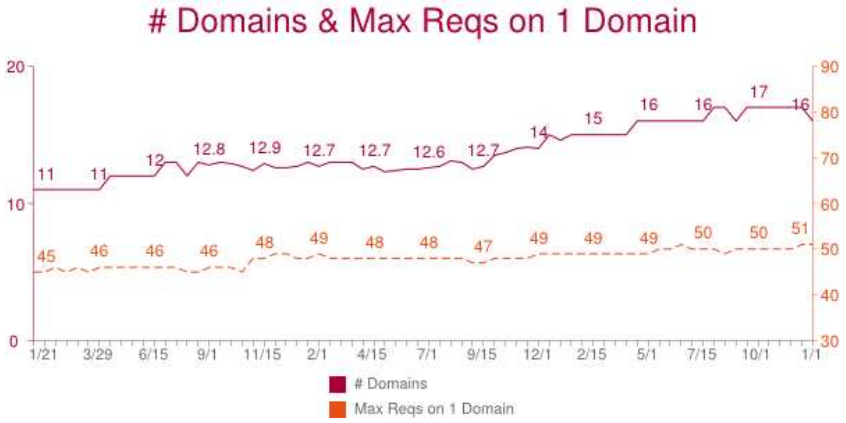


Figure 3.1: Evolution of cross-domain requests between January 2011 and January 2014, data and chart by `httparchive.org`

Figure 3.1 shows the evolution of cross-domain requests between January 2011 and January 2014.[4] It is based on data gathered by the HTTP Archive, which is part of the Internet Archive consortium, on all its monitored websites. A single web page typically loads resources from a variety of web servers across many domains. The chart shows first the evolution in the average number of domains that are accessed across all websites, and second the average of the number of requests on the most-used domain.

Thus, figure 3.1 highlights that in recent years the number of external domains used to fetch content on every single page has been constantly growing. By 2014, it is common for the content of a web page to be put together from no less than 16 external domains (i.e. external websites) on average.

Additionally, the most popular content providers are always more solicited. Concrete examples include CDNs (content delivery networks) which are dedicated to images, page layout (CSS), or videos. External scripts libraries are another example, as well as actual text data streams from multiple sources.

As a result, so-called “dynamic pages" may change anytime without prior notice, simply because the embedded content has changed. The practical shortcomings of the evolution towards the “Web 2.0" (an expression usually used to refer to increasingly dynamic web pages) is twofold. The requested content may already have disappeared or changed between query and actual download, or it may even change faster than the search engines can index it.

Conceptual issues Besides, the last decade has seen a slow but inescapable shift from “web as corpus" to “web for corpus" due to an increasing number of web pages and the necessity of using sampling methods at some stage. In that sense, the question whether the method used so far, that is randomizing keywords, provides a good overview of a language is still open. It now seems reasonable to look for alternatives, so that research material does not depend on a single data source, as this kind of black box effect combined with paid queries really impedes reproducibility of research.

“The assumption is that sending 3- or 4-tuples of mid-frequency terms as conjunct

[4] `http://httparchive.org/trends.php?s=All&minlabel=Dec+28+2010&maxlabel=Jan+1+2014#numDomains&maxDomainReqs`

> queries to search engines and harvesting the first n results is a random enough process or at least that it delivers documents which are relevant and diverse enough for corpus construction. It has to be kept in mind, however, that the returned documents will at least be biased towards word lists and long documents. Whether a corpus designer is satisfied with this is an individual decision." (Schäfer & Bildhauer, 2013, p. 18)

By querying random tuples, one may find web pages which for one reason or another publish randomly aggregated content. This is not a mere theoretical hypothesis, since as said by Schäfer and Bildhauer (2013) in the quote above, there are many word lists on the web, sometimes logically among the very first links found when using a randomly generated query.

Regarding the length of documents mentioned by Schäfer and Bildhauer (2013), research practice has indeed shown that search engines usually tend to favor pages with more text. A detailed analysis of this phenomenon is performed in the next chapter below. While it may seem to be an immediate advantage for web corpus building, it also entails a strong bias towards long texts, which is not always desirable.

For example, a blog containing witty aphorisms, despite its quality, may have to optimize its content in order to be found easily by search engines, e.g. by publishing many posts per web page. On the contrary, a blog with very long, but potentially repetitive, posts could score higher in the search engines' ranking, not only because of the nature of the texts but simply because there are more words to be found and to be used as keywords. However, from the perspective of a linguist, the aphorism genre is probably as important as the "very long blog posts" genre. That is why the bias in search engines is significant.

Alternatives The Clueweb part of the Lemur Project[5] has a graph-based approach to web knowledge collected from previous crawls. In Clueweb's case, graph theory could allow for the discovery of more productive start URLs. In the last version of the project, URLs coming from Twitter are also classified using graphs.

The reasons for the graph-based approach emphasize that is it crucial to offer solutions for issues of seed productivity and the increasing role of social networks. Concrete analyses concerning the latter are detailed in the following chapter below, which also deals with the importance of having "better" or more efficient seeds.

Through its attention to seeds and its advanced spam filtering techniques, the Clueweb project, led by computer scientists, seems to be an interesting project. However, information and documentation about the methods used is scarce.

3.2.3 In the civilized world: "Restricted retrieval"

3.2.3.1 Definition

What I choose to name restricted or circumscribed retrieval here is sometimes called "focused" or "scoped" crawling in the literature. However, the interpretation of these terms may vary since they are also used in a web crawling context (see crawling strategies p. 121).

I define restricted retrieval as the download of resources where the location is already known, or at least supposed. A straightforward way to perform this operation is to use a

[5] http://boston.lti.cs.cmu.edu/clueweb12/

list of URLs, identifying resources on the web, or to download the archive of a website when available. As such there are no unknowns and generally no surprises concerning the content.

3.2.3.2 Implementation

Single source A first example is for instance a newspaper website where the number and the location of articles to download are already known, be it because all the available links are retrieved, for instance using a sitemap, or because of prior selection by hand or using URL patterns like a column name or a date. Details of the corpus building process from newspaper websites are given in the next chapter below.

So-called Wikipedia dumps[6] are another option for gathering text from a single source. The dumps are bundled archives of Wikipedia or other Wikimedia projects like the Wiktionary for a given language. They are retrieved using only one link and downloading relatively large archive files which once extracted contain part of or all of the information available on the project in a text or database format. Even if this resource is well-categorized and furnished with relatively clean metadata, it still has to be processed in order to become a text corpus in a linguistic sense, so that the steps described above are roughly similar to other web resources.

Multiple sources Circumscribed retrieval can also be performed from multiple sources, be it multiple websites from which the pages are downloaded or multiple input sources. The use of an URL directory is an example of the first case. The human-edited directory DMOZ[7] is a carefully curated resource where links can be found according to precise categories, for instance weather forecasts in a given language or pages dealing with the rules and regulations of cricket. While the directory is a single website, it links to a lot of different sources, currently more than 4 million websites. Using it as an input may lead to a much greater variety of information sources, for example in order to build topic-based corpora.

The BootCaT method (Baroni & Bernardini, 2004) described above (see p. 116) is an example of the latter case, multiple input sources. It consists in querying search engine with random words combinations in a given language in order to gather URLs of documents to retrieve. In that way, and if only these URLs are used, i.e. without crawling, it is a finite list which is processed at once, and this a restricted collection. The advantages and limits of this method are described in detail below.

Possible extensions of the concept Possible extensions of the notion include the traversal (crawling) of a single website in order to find pages which may not be included in the original list. These could for example be new, recently published articles. In that sense, the crawling of a particular website falls under the category of restricted retrieval (for a definition of crawling see p. 111).

Multiple restricted retrievals are also a possible extension, in order to build a corpus from several sources. The Czech Internet corpus described in the previous chapter (see p. 52) may still be called a restricted approach, although it makes use of web crawling techniques. Indeed, the corpus has been compiled using “particular, carefully selected sites" (Spoustová & Spousta, 2012, p. 312), which were known in advance. However, it is not clear whether precise lists of URLs were established prior to a massive download or if part of the web pages were discovered

[6] http://dumps.wikimedia.org
[7] http://www.dmoz.org

on-the-fly. This makes it a borderline case which in a broad sense still belongs among the circumscribed retrieval approaches.

In sum, this kind of document retrieval may imply techniques close to web crawling but is not web crawling in the common sense of the word, since at best a website is traversed and not the web itself. The retrieval of URLs from a directory for instance would cease to be restricted if the links are followed, leading to documents which were not in the initial list.

3.2.4 Into the wild: Web crawling

3.2.4.1 Too complex a task for linguists?

Contrary to a retrieval where the URLs are known in advance, web crawling implies the discovery of new resources during the course of a web traversal. It is a document retrieval in the open, which can take the form of a large web space exploration but is not necessary as broad and as long as possible. The portion of the web which is to be taken into account can be restricted in several ways, to be described below. It still is significantly larger or deeper than in the methods described above.

Per se, web crawling may yield inconstant results, as the exploration course is not always fixed in advance and it does not systematically go as planned. This change from a swimming-pool to the open sea is a quantitative and qualitative jump, for which a much higher degree of robustness is needed as well as a knowledge of certain web usage rules.

As the web in itself or the potential target of the study expands, the complexity of the crawling task increases. It may require a lot of machine power as well as a recent, well-designed software environment:

> “A reasonable rate can only be achieved by a multi-threaded program or a fully parallelized piece of software capable to run on a cluster of machines. This, however, increases the complexity of the whole system, since the huge data structures [...] need to be synchronized between the different machines." (Schäfer & Bildhauer, 2013, p. 20)

Thus, it is a more difficult challenge, which is a fact I would like to highlight, since web corpus construction ought to be mastered from beginning to the end. It may be acceptable to term web crawling processes a technical world for practical reasons having to do with work division. But I believe that it is not acceptable in any way for linguists to call it a mere engineering problem and to totally leave the field to “crawl engineers" and the like, because it has deep consequences on research conditions and results.

3.2.4.2 Crawling steps

Schäfer and Bildhauer (2013, p. 16) identifies four different steps in a routine crawl, which I reformulate as follows:

- Collect seed URLs (the URLs needed to start the crawl)
- Decide what kind of content to target and implement subsequent constraints (URL-based restrictions, MIME types, file size, language, encoding, etc.)
- Take a stand on politeness settings

- Run the crawler and eventually monitor the course of the crawl

None of these steps can be considered trivial, they all have a sometimes underestimated impact on the final result. The seed URLs are the starting points in the web graph, and they have an impact on the beginning of the crawl at the very least (see chapter 4 for the results).

The content restrictions ought to be carefully implemented; on the one hand, from a recall-based perspective as well as in order to avoid missing data that could be useful because of unexpected and/or tricky URL structures, MIME, or encoding declarations. On the other hand, precision can be crucial with large corpora due to an increasing processing time. Efficiently filtering unwanted website types (e.g. video platforms) and/or content types (e.g. adult content) can avoid a dramatic waste of bandwidth and processing time.

Politeness settings depend on a whole abuse response infrastructure. The longer the crawl, the most probable it is that the crawler will be blocked and that network administrators will be contacted. Politeness of crawlers is discussed more in detail below (see p. 125)

Last, detecting and acknowledging potential software failures or undesirable crawling courses can save a lot of time and energy.

3.2.4.3 Crawling strategies

Breadth-first search (BFS)

> "The basic web crawling algorithm is simple: Given a set of seed Uniform Resource Locators (URLs), a crawler downloads all the web pages addressed by the URLs, extracts the hyperlinks contained in the pages, and iteratively downloads the web pages addressed by these hyperlinks." (Olston & Najork, 2010, p. 178)

BFS is believed to be a simple crawling strategy that is frequently used, including in web linguistics projects, for instance by Baroni et al. (2009).[8] It consists in extracting all the links of the downloaded web pages and visiting them one by one, the only restriction being that a single URL should not be visited more than once.

> "The simplest and most widely used crawling strategy [...] is breadth-first, a traversal technique. In a pure breadth-first crawl, all links are harvested and enqueued in the order they are found." (Schäfer & Bildhauer, 2013, p. 28)

Breadth-first search is an efficient technique: since all the links are gathered and processed, it can be a convenient way to collect a substantial number of URLs in a short period of time. However, the seed URLs have a strong influence on the course of a crawl, because the very first web pages downloaded can determine the general direction of the crawl.

> "This strategy leads to the exhaustive collection of a local connected sub-component of the web graph. BFS can thus introduce a bias toward the local neighborhood of the seed URLs, possibly not discovering relevant/interesting material in other parts of the graph." (Biemann et al., 2013, p. 25)

[8] "The crawls are performed using the Heritrix crawler, with a multi-threaded breadth-first crawling strategy; they are stopped after 10 days of continuous running." (Baroni et al., 2009, p. 214)

This strategy is also heavy on server-side: if the pages of a given website are closely interlinked, the server which hosts the site will receive many more requests, which can lead to a bulky management of queries by server or IP, depending on the politeness settings chosen.

Because of these limitations, other strategies have been developed.

Scoped crawling In its general definition, scoped crawling is a way to make crawls more efficient by using constraints on either content or retrieval:

> "Originally, crawlers optimized to find documents according to genres, languages, etc. are called scoped crawlers" (Schäfer & Bildhauer, 2013, p. 34)

What seems to define scoped crawling is indeed the intent to retrieve a particular section of the Web, which is made possible by a series of constraints implemented in the crawling software:

> "Scoped crawling imposes constraints on the kind of crawled documents. The idea is to avoid downloading content which would not be included in the final corpus anyway. A scoped crawl is restricted by accepting only documents from certain URLs, IP ranges, etc." (Biemann et al., 2013, p. 26)

So, URLs or IP ranges have possible tangible characteristics on which constraints may be applied. A main reason for doing this is a greater efficiency of the whole process. In fact, one may think that the desired content would be acquired anyway, regardless of the crawling strategy and the constraints, provided a crawl is long and/or extensive enough. However, due to the size of the Web, this is not a realistic option, for instance for less-resourced languages, which represent only a tiny fraction of the available websites (Scannell, 2007).

There are also finer criteria, such as language, topic, or location, as Olston and Najork (2010) explain:

> "A scoped crawler strives to limit crawling activities to pages that fall within a particular category or scope, thereby acquiring in-scope content much faster and more cheaply than via a comprehensive crawl. Scope may be defined according to topic (e.g., pages about aviation), geography (e.g., pages about locations in and around Oldenburg, Germany), format (e.g., images and multimedia), genre (e.g., course syllabi), language (e.g., pages in Portuguese), or other aspects." (Olston & Najork, 2010, p. 208)

A standard way of performing scoped crawling is to target a particular top-level domain (TLD) to try to maximize the number of pages in a given language. Choosing a TLD like `.dk` for instance could lead to more pages in Danish than any other one.

However, due to the popularity of certain TLDs (such as `.me` for Montenegro or `.tv` for Tuvalu) and the increasing number of available websites, TLDs and websites' country of origin are not as closely correlated as in the beginning of the 2000s. Additionally, high-density languages are not necessarily predictably distributed among major TLDs (let alone the domains in `.net` or `.com` which are not linked to a particular language), while low-density languages may simply not be available under a particular TLD.

Focused crawling There is no clear divide between focused and scoped crawling. The first can be defined as a narrower version of the latter, an "intelligent" crawling strategy.

> "Topical crawling (also known as "focused crawling"), in which in-scope pages are ones that are relevant to a particular topic or set of topics, is by far the most extensively studied form of scoped crawling." (Olston & Najork, 2010, p. 209)

In fact, constraints are more specific in the case of focused crawling. The distinction could be twofold. First, they are expected to be finer, for example they do not necessarily rely on domain information. Second, they involve the implementation of more developed decisions processes, such as heuristics based on URLs or information gathered in the course of the crawl.

> "Focused crawling imposes even stricter constraints and tries to efficiently discover specific types of information (languages, genres, topics, etc.) which cannot simply be inferred from address ranges, domains, or hosts. A focused crawler guesses for each harvested link the kind of document it points to." (Biemann et al., 2013, p. 27)

Prediction is an important dimension of this strategy. It implies that the crawler takes chances depending on fine-tuned settings which may vary in the course of time. The definition of "link usefulness" may also change during a crawling project.

> "Usually, the links themselves within a document and the text in a window around the links are analyzed to come up with a prediction about the usefulness of the link given the crawl's scope or focus. Often, the relevance of the linking page itself or a larger link context are also included in the calculation. Various heuristics and machine learning methods are used for the calculation." (Schäfer & Bildhauer, 2013, p. 34)

In sum, while it can be expected that breadth-first search and scoped crawls are reproducible provided the seed URLs are the same, focused crawls may not lead to the same pages, for instance because the content of the web pages changed, thus altering the calculations.

Optimized crawling Last, a subcategory of focused crawling is sometimes called optimized crawling. It mostly refers to the technical efficiency of the crawl:

> "What we call optimized crawling is similar to focused crawling, but it is merely intended to make the crawl more effective, so that more usable text documents can be downloaded in a shorter time, wasting less bandwidth. [...] The crawl is biased towards documents which, according to some metric, have a high relevance." (Biemann et al., 2013, p. 27)

Vertical crawling Last, the notion of vertical crawling is introduced by Olston and Najork (2010). It is described as a specialization on the side of the crawler regarding a particular website or a particular type of content, due to its importance:

> "Some sites may be considered important enough to merit crawler specialization [...]. Also, as the web matures, certain content dissemination structures become relatively standardized, e.g., news, blogs, tutorials, and discussion forums." (Olston & Najork, 2010, p. 237)

The purpose of such a crawler is increased efficiency, as in focused and optimized crawling:

> "A crawler that understands these formats can crawl them more efficiently and effectively than a general-purpose crawler." (Olston & Najork, 2010, p. 237)

The various specialized crawler types described here are not to be considered strictly separate, there are complementary approaches, and the differences can be subtle. According to Olston and Najork (2010), vertical crawling focuses on syntactic rather than semantic handling.

> "Scoped crawling focuses on collecting semantically coherent content from many sites, whereas vertical crawling exploits syntactic patterns on particular sites." (Olston & Najork, 2010, p. 238)

Thus, a vertical crawler is one that addresses a particular type of markup, based on surface factors, without discarding any kind of content.

The notions of optimized crawling are not be used further in this document. The next chapter below focuses on BFS, and scoped as well as focused crawling, with hints on vertical crawling.

3.2.4.4 Limitations of web crawling

Limitations of web crawling include the size of the web, inaccessible documents, URL issues and the fast-paced evolution of the web.

Completeness vs performance First of all, the web is far too large to be crawled completely, even by the most powerful search engines. Thus, crawls are directed in a way full of approximations and sampling processes, where even a clear crawling strategy may lead to a biased result, because the text collection at the end of the download is at best a tiny subpart of the web.

> "Building corpora from the web starts by sampling from a population and will usually result in a corpus that is orders of magnitude smaller than the web itself." (Biemann et al., 2013, p. 24)

It is interesting to see the sampling problem of the first electronic corpora being revisited here in another context, at another scale, with other technological obstacles and other consequences.

> "Even the highest-throughput crawlers do not purport to crawl the whole web, or keep up with all the changes. Instead, crawling is performed selectively and in a carefully controlled order. The goals are to acquire high-value content quickly, ensure eventual coverage of all reasonable content, and bypass low-quality, irrelevant, redundant, and malicious content." (Olston & Najork, 2010, p. 178)

“Deep web" Second, there are many web pages which could be linguistically interesting but cannot be crawled because they require user authentication, some sort of user feedback, or a complete browser-like software environment for instance. Such pages are called the “deep web", with contours that can at best be sketched but whose contents remain mostly inaccessible despite mitigation techniques.

> “A news archive contains content that is primarily unstructured (of course, some structure is present, e.g., title, date, author). In conjunction with a simple textual search interface, a news archive constitutes an example of an unstructured-content/unstructured-query deep web site." (Olston & Najork, 2010, p. 231)

“Politeness" of robots

> “Crawlers should be “good citizens” of the web, i.e., not impose too much of a burden on the web sites they crawl." (Olston & Najork, 2010, p. 178)

In fact, the `robots.txt` guidelines (Koster, 1994), describes a way for site operators to restrict crawler coverage on their website. They are not officially standard but have become a largely followed and thereby de facto standardized approach.

> “Web crawlers are supposed to adhere to the Robots Exclusion Protocol, a convention that allows a web site administrator to bar web crawlers from crawling their site, or some pages within the site. This is done by providing a file at URL /robots.txt containing rules that specify which pages the crawler is allowed to download. Before attempting to crawl a site, a crawler should check whether the site supplies a /robots.txt file, and if so, adhere to its rules." (Olston & Najork, 2010, p. 191)

If followed, robot exclusion can impose strong restrictions on corpus construction. If a particular website is targeted and if the `robots.txt` instructions rule out the crawling robot, then it is ethically not recommended to perform the crawl, as it is comparable to gathering of information without consent.

Web pages may also be inaccessible because they block crawling robots one way or another. Although these pages are not technically part of the deep web, the effect is the same, with a major difference: the decisions in that case mostly rely on crawling policies and not on practical obstacles.

Ever-changing status of URLs and documents behind them Content personalization is a characteristic feature of the move towards so-called “Web 2.0". The personalized user experience is supposed to enhance websites, for example in order to make them more enjoyable, and thus more popular, as one does not see all the possible kinds of content the site has to offer, only what is expected to be interesting relative to one’s prior behavior on the website and on the Internet in general.

The main draw of personalization may well have to do with the way websites are monetized. In fact, it is far more efficient in terms of clicks to perform user profiling so that ads shown on a given page correspond one’s potential centers of interest. Whether they are aware of it or not, nearly every Internet user does not surf the Web anonymously, due to the current features of websites it is relatively easy to follow a given user no matter where he is.

The same is true for robots. If the crawls are large and noticeable enough, they are bound to lead to particular reactions by the websites. A malicious example would be a website which “hides" its real nature in front of Googlebot (let us say the fact that every single page is full of ads) but which systematically serves its full content to users others than robots.

A more useful example is the case of content language. If the machine located in Russia and serving a web page detecting a user coming from Germany, and if it can then serve content in German rather than in Russian or in English, it is logical to think that this might be a better option. Thus, the website will be in German, and there will probably be a menu for the user to change the language if desired. Since most robots are not “intelligent" enough to do that, the crawler will end up with a page in German, even if the purpose was to fetch a page in Russian because it was known that the server was probably in Russia.

> “A feature of the Web is that, depending on the details of a request, different representations may be served up to different requesters. For example, the HTML produced may vary based on conditions hidden from the client (such as which particular machines in a back-end server farm process the request) and by the server’s customization of the response. Cookies, representing previous state, may also be used, causing different users to see different content (and thus have different links in the Web graph) based on earlier behavior and visits to the same or to other sites. This sort of user-dependent state is not directly accounted for in current Web-graph models." (Hendler et al., 2008, p. 64)

As Hendler et al. (2008) explains, the Janus-faced behavior of certain web servers (technically termed as cloaking) does not only impact a single web page, but also the series of links it contains. As a consequence, it can affect the whole crawl. I personally experienced a crawl designed to target Malaysian web pages ending up containing more than 80% of web documents in German (see chapter 4 for a discussion).

To conclude this point, because of a combination of factors are related to user profiling and user experience personalization mostly falling under the categories of cloaking and Web 2.0, the Web is bound to change rapidly and, according to the requester’s point of view:

> “The Web is different from most previously studied systems in that it is changing at a rate that may be of the same order as, or perhaps greater than, even the most knowledgeable researcher’s ability to observe it." (Hendler et al., 2008, p. 68)

Additionally, as content on the Web also ages very quickly, “old" or invalid URLs are also a major difficulty. On the one hand, links are ephemeral and so is the content behind them, so that “not even two exactly synchronous accesses to a web page necessarily lead to the same document" (Schäfer & Bildhauer, 2013, p. 10). On the other hand, a URL can take many forms depending on the publishing system, including incomplete or invalid addresses, even if the so-called canonical form of a URL is standardized. Issues regarding URLs are described in further detail in chapter 4.

3.2.4.5 The open issue of web representativeness

Concerning representativeness seen from a more technical perspective, the same authors (Berners-Lee et al., 2006) mention usage of randomness and sampling processes as possible cues, but even that is difficult to assess with precision.

> "How should a sample be gathered in order to be properly called representative? To be properly useful, a sample should be random; 'randomness' is usually defined for particular domains, and in general means that all individuals in the domain have an equal probability of being selected for the sample. But for the Web that entails, for example, understanding what the individuals are; for instance, are we concerned with websites or web pages? If the former, then one can imagine difficulties as there is no complete enumeration of them. And sampling methods based on, say, IP addresses are complicated by the necessarily sparse population of the address space. Furthermore, so cheap are operations on the Web that a small number of operators could skew results however carefully the sample is chosen." (Berners-Lee et al., 2006, p. 13)

Thus, sampling methods are not strictly impossible, but they are technically complex, and can be tricked easily. All in all, this situation explains why the notion of representativeness is challenged by web-scale research, and why restrictions are necessary in order to make corpus construction possible on a simpler basis. In fact, the priorities shift towards goals easier to reach:

> "The rationale here is for the corpus to include a sample of pages that are representative of the language of interest, rather than getting a random sample of web pages representative of the language of the web. While the latter is a legitimate object for 'web linguistics', its pursuit is not among the priorities set out for the WaCky corpora." (Baroni et al., 2009, p. 213)

So far, web corpora have been built without investigating the question of web representativeness any further, although researchers acknowledge that it may be a topic of interest for linguists truly committed to web data.

Moreover, the hypothesis that existing web crawling techniques necessarily impact the content one way or another is also considered to be a valid subject of inquiry.

> "The questions of how more advanced sampling methods affect the final Web corpus, and how the size of the final corpus given a certain crawling effort can and should be balanced against the degree of validity of the sampling procedure have not been examined for linguistically oriented Web corpus construction." (Schäfer & Bildhauer, 2013, p. 34)

Knowing more about web representativeness can enable researchers to redefine and adapt the notions of balance and representativeness to web corpora. However, there is still much to do on this topic, which I think is worth mentioning but which due to its scale is not part of the work presented in this thesis.

3.2.5 Finding a vaguely defined needle in a haystack: targeting small, partly unknown fractions of the web

3.2.5.1 Less-resourced languages

Main challenges Concerning less-resourced languages, many methodological issues remain leading to different notions of web corpora and different expectations towards the experimental reality they offer.

A major issue is precisely the lack of interest and project financing when dealing with certain low-resource languages, which makes it necessary to use light-weight approaches where costs are lowered as much as possible:

> "The lack of funding opportunities or commercial interest in this work has led to an approach based on certain principles that offer maximal 'bang for the buck': monolingual and parallel corpora harvested by web-crawling, language-independent tools when possible, an open-source development model that leverages volunteer labor by language enthusiasts, and unsupervised machine learning algorithms." (Scannell, 2007, p. 5)

Another issue resides in the potential sources, which have to be found and properly evaluated:

> "The first major challenge facing any corpus builder is the identification of suitable sources of corpus data. Design criteria for large corpora are of little use if no repositories of electronic text can be found with which to economically construct a corpus." (Baker et al., 2004, p. 510f)

No consensus in research practice There is no consensus to be found among the existing techniques. URL classification problems for instance make a proper language identification of the content necessary: especially for lesser-known languages, it is not so easy to find working patterns like those used by Baykan et al. (2008), who try to classify web pages as to their main language only by examining the web pages' URLs.

If it is not possible to determine the nature of the content without seeing it, the way web documents are classified on the crawling side is not clear either.

Baker et al. (2004) state that it was faster for them not to use any automatic crawling and to turn to hand-picked content:

> "We found it was faster for a human to visit the site, sort the text from the adverts, identify the useful material and save it." (Baker et al., 2004, p. 511)

However, the technical state of the art is more developed now than it was in 2004. In the previous years, at least two major projects relied on crawling techniques only and not hand-picked content anymore. The Crúbadán[9] project was originally devoted to the study of Celtic languages, but was later adapted to target several hundreds of minority languages. The project researchers chose to focus on one language at a time because crawling the whole web was considered a waste of time and resources:

> "The crawler focuses on one language at a time. A reasonable alternative would have been to crawl the web very broadly and categorize each downloaded document using the language recognizer, but this is clearly inefficient if one cares primarily about finding texts in languages that do not have a large presence on the web." (Scannell, 2007, p. 10)

[9] Literally "crawler" in Irish, but with the additional (appropriate in this context) connotation of unwanted or clumsy "pawing", from the root crúb ("paw").
http://borel.slu.edu/crubadan/

On the other hand, the Leipzig Corpora Collection (Goldhahn, Eckart, & Quasthoff, 2012), which started as the FindLinks project (Heyer & Quasthoff, 2004), is an example of a global approach, but little is known about the crawling methods used, other than their being breadth-first, starting from a directory of news websites as source for many less-spoken languages, which seems to be successful.

Then again, Scannell (2007) states that crawling without expert knowledge is "doomed to failure", which shows that there is no consensus on this point either.

Problems and benefits of document nature The frequency of mixed-language documents is unevenly distributed on the web. As a matter of fact, they have been shown to be more of a problem concerning minority languages, making it even harder to gather web texts in these cases:

> "The majority of web pages that contain text in a minority language also contain text in other languages." (King & Abney, 2013, p. 1110)

As for the reasons for using several languages in a single web document, code-switching alone does not seem to be a convincing approach. meaning that the study of mixed-language documents for linguistics purposes is probably irrelevant:

> "Though code-switching has been well-studied linguistically, it is only one possible reason to explain why a document contains multiple languages" (King & Abney, 2013, p. 1111)

Thus, it can be said that the texts are different in nature, which may have benefits in some cases, for instance concerning text quality and more precisely machine-generated text:

> "One benefit of working with under-resourced languages is that they are only rarely the target of 'WAC spam' – documents not written by humans who speak the target language but instead generated automatically by a computer one way or another." (Scannell, 2007, p. 13)

Conclusion: more downsides than advantages To conclude, one may say that the study of less-resourced languages on the Web bears overall more downsides than advantages. Indeed, easiness factors, such as the lack of spam and the smallish community of users, are outweighed by factors of complexity, such as the difficulty of the crawling process and the extrinsic and intrinsic variation of documents, and last but not least the difficulty to fund research projects on this topic.

3.2.5.2 Computer-mediated communication and afferent specific text genres

So far, there are few projects dealing with computer-mediated communication (CMC), mostly due to the recent expansion of social networks and the fast pace at which those resources are changing. Most scientific studies are focused on information extraction, such as sentiment extraction, e.g. whether tweets on a particular topic are rather positive or negative, or comparisons based on geographic information, e.g. by using the metadata of short messages.

In the case of German, the DeRiK project features ongoing work with the purpose of building a reference corpus dedicated to CMC (Beißwenger et al., 2013), which could be used in

order to study German and its varieties as they are spoken on the Web. More specifically, this kind of corpus can be used to find relevant examples for lexicography and dictionary building projects, and/or to test linguistic annotation chains for robustness.

The problems to be solved in order to be able to build reliable CMC corpora are closely related to the ones encountered when dealing with general web corpora as described above. Specific issues are threefold. First, what is relevant content and where is it to be found? Second, how can information extraction issues be tackled? Last, is it possible to get a reasonable image of the result in terms of text quality and diversity?

3.2.5.3 Summary of issues to address

Targeting small, partly unknown fractions of the web enables researchers to target different speaker communities and various text types. It also involves dealing with potentially extreme content variability, and consequently requires adaptation of the crawling strategies.

Because the objects of study are still new, such as in the case of social networks, or difficult to fund, as in the case of less-resourced languages, there is no consensus in research practice. There is also a real potential in terms of discoveries, provided the researchers manage to maximize the efficiency of the crawls and to provide the most “bang for the buck" (Scannell, 2007).

Rich metadata considered to be a desideratum for web corpora by Tanguy (2013), are particularly relevant in that case, be it for language documentation purposes or to do justice to the variability of the texts and microtexts which have been gathered.

The issue of less-resourced languages is addressed in section 4.2.1.2, while microtexts, social networks, and computer-mediated communication are studied in section 4.2.1.3 concerning a cross-language comparison.

3.3 An underestimated step: Preprocessing

3.3.1 Introduction

3.3.1.1 Why “pre-" rather than post-processing and why is it important?

The designation of pre-processing I choose to name the operations described in this chapter pre- and not post-processing, as it is frequently named in the literature. This designation is also used by Eckart, Quasthoff, and Goldhahn (2012), while Baroni et al. (2009, p. 214) is more specific by speaking of “post-crawl cleaning".

First of all, a lot of text is discarded (see table p. 138), so that the word post-processing is misleading.

Second, there are a number of steps where the chosen software architecture is decisive regarding the quality of the end product.

Last, the shift from what many researchers call “post-processing", relative to the crawl and the download, to “pre-processing", relative to the annotation and indexing toolchain making the corpus actually accessible, illustrates the necessity to put the corpus back into focus, since “pre-" clearly indicates that the operations determine the quality of the final corpus.

Major steps A definition of pre-processing, i.e. in their terminology post-processing, is given by Schäfer and Bildhauer (2013):

> "By post-processing, we refer to the non-linguistic cleanups which are required to turn the collection of downloaded HTML documents into a collection of documents ready to be included in a corpus, which involves cleanups within the documents and the removal of documents which do not meet certain criteria. The end product of the post-processing chain is often simply a set of plain text documents." (Schäfer & Bildhauer, 2013, p. 36)

The major steps required to process downloaded web pages are known, the description of Lüdeling et al. (2007) is still accurate:

> "Once a set of web pages has been crawled and retrieved, one has to strip off the HTML and other 'boilerplate'. The character encoding and language of each page must be identified. 'Linguistically uninteresting' pages (e.g. catalogues and link lists) must be discarded. Identical and – much more difficult – 'nearly identical' pages have to be identified and discarded." (Lüdeling et al., 2007, p. 19)

However, their statement that "none of these tasks is particularly difficult per se" can be put into question.

Example In the following first example, the preprocessing chopped the markup correctly and "revealed" a useful sentence with atypical word forms and syntax in the perspective of "classical" (non-web) corpora:

```
wenn ich könnte würd ich nen bot schreiben , der an alle in einer liste
on the fly noFollow anhängt und dann erst recht diese komische komponente
auf nem nichtsnutzigen webspace online stellen , natürlich ohne jedwede
einnahmen durch den kram .
```

DECOW2012X – slice 3, 42631536[10]

The following sentence exemplifies a problem in boilerplate removal which apparently truncates an existing valid sentence, which greatly affects qualitative analysis:

```
Eine weitere Quelle wird ein rel = " nofollow " onclick = " \_gaq.push
( [ '\_trackPageview ' , '/ outgoing/ article\_exit\_link/ 4017961 ' ] ) ;
" = javascript " wikipedia .
```

DECOW2012X – slice 3, 501147766[11]

Both examples make clear that preprocessing is not a trivial task. In fact, it is even a complex one, involving a series of different steps.

3.3.1.2 Filtering steps

According to Eckart et al. (2012), there are two main filtering steps, HTML-stripping and text cleaning, the latter being divided into five different steps:

[10] (Schäfer & Bildhauer, 2012)
Available through the COLiBrI interface `http://hpsg.fu-berlin.de/cow/colibri/`

[11] *Ibid.*

1. HTML-Stripping (Remove all HTML-code and additional markup)
2. Text cleaning
    a) Boilerplate removal (see p. 133 for a definition)
    b) Removal of foreign language parts (whole texts or sentences)
    c) (Optional) Removal of parts which are not well-formed sentences, using pattern matching methods
    d) Removal of duplicate sentences
    e) Removal of near duplicate sentences

The steps set out above concern “usual" IE/IR software as well as software such as used by commercial search engines, but in order to build web corpora it features a touch of natural language processing. The whole process has a huge potential in terms of impact, as described below (see p. 138).

3.3.2 Description and discussion of salient preprocessing steps

3.3.2.1 Web document rejection based on simple characteristics

Most of the time, web characteristics such as the size or the MIME-type[12] are used to automatically discard pages bound to be not of interest. These characteristics are present in the HTTP header which is fetched according to the HTTP protocol, i.e. before the download of the proper content. The usual size limit is set at a few megabytes, while the accepted MIME types include text and HTML files and sometimes PDF for instance.

As an example of how crawling processes need to be robust, it is occasionally necessary to cut the download at a given point because the web page has reached a certain size due to undesirable content features such as never-ending feed, or video streaming.

Moreover, the size of downloaded files is also relevant for making guesses regarding their content. As such, size can be used to discard web documents:

> “Very small documents tend to contain little genuine text (5 KB counts as ‘very small’ because of the HTML code overhead) and very large documents tend to be lists of various sorts, such as library indices, store catalogs, etc." (Baroni et al., 2009, p. 214)

Thus, monitoring the downloads and reacting (automatically or manually) to events occurring during this step is desirable to ensure that the crawler is kept in good working order, but also to start a first filtering which makes the archives lighter.

3.3.2.2 Markup and encoding cleanup

Further down the processing chain, the content of the downloaded files has to be scanned by robust tools in order to take necessary action. The first issue resides with faulty markup, which may not be visible to end users:

[12] MIME is a content type convention.

> "HTML as found on the Web does often not conform to the expected standards. This is mostly because Web browsers are quite fault-tolerant when rendering HTML, such that errors in the markup are not visible." (Schäfer & Bildhauer, 2013, p. 39)

Potential errors in the markup can result in non- or malfunctioning tools if they are not detected and repaired properly. Since HTML irregularities are relatively frequent, for example due to embedding of content from other sources (see p. 117), a whole range of tools exist whose purpose is to clean and/or tidy the HTML code of a particular web page.

Secondly, web page encoding can be a tricky issue, particularly as one tackles unknown documents extracted from the Web. In fact, the main problem is that the genesis of web documents is not known, all that one can do is guess regarding their encoding history, using hints such as the advertised encoding, when available, or the de facto algorithm of the content. However, hints gathered this way do not necessarily match, sometimes for a very good reason, for example the presence of several encodings on a single web page:

> "For many languages, Web documents come in more than one encoding." (Schäfer & Bildhauer, 2013, p. 40)

Depending on the target language, usual cases may be identified and remedied with a satisfying accuracy. In the case of German for instance, irregularities can stem from a erroneous encoding recognition which leads to an improper labeling, for instance a text published using an older version of Windows (CP-1252), but improperly labeled as the more recent Latin-1 encoding (ISO-8859-1), since both encodings share a large majority of characteristics.

Although well-known or frequent problems may be isolated and corrected, markup and encoding cleanup remains a tough challenge that requires downstream tools to be robust in order not to break on HTML or encoding irregularities. This issue is particularly present when dealing with web texts in numerous languages and generated by countless publishing systems. There are no less than four different encodings used for Japanese, and unclean HTML can be found in every corner of the Web. These issues are not trivial, since they contribute to the reign of the unpredictable in the corpus construction process.

3.3.2.3 Boilerplate removal

Definition If clean markup and consistent encoding seem abstract, although they truly impact the end corpus, boilerplate removal is the operation which is probably the most salient in the way web corpora are experienced by linguists.

Boilerplate removal takes place after markup removal, and it involves deleting unwanted parts which undermine text cohesion.

> "After a successful removal of the markup, only the text (without the graphics and other multimedia content) that would be visible in a version of the page as rendered in a Web browser remains. This text, however, might contain parts that we do not want to include in a corpus." (Schäfer & Bildhauer, 2013, p. 47)

Approaches There are two main approaches to boilerplate removal, on one side HTML-based ones and on the other statistical ones.

Many different features can be used, which makes the task even harder: markup-related features; stop words; graphemic features, such as numbers or upper case use; linguistically motivated features; and whole-document features, such as length or ratios (Schäfer & Bildhauer, 2013, p. 50f).

HTML-based approaches Html2Text is a tool which is available on UNIX systems[13]. It relies on series of simple filtering rules designed so that the result looks like a copy-paste of browser rendering. There is no element hierarchy in the final document, which is supposed to keep each and every word available in the original document.

Lynx is a text-based web browser. Browsing in Lynx consists of highlighting the chosen link using cursor keys, or having all links on a page numbered and entering the chosen link's number. The formatting provided by Lynx can be used to extract relevant information in a page.

Approaches using statistical patterns and/or machine learning Machine learning plays a decisive role in automatic processing. The granularity as well as the decision factors have to be set in advance, then the learning is done by a training phase on known, labeled data.

> "Automatic boilerplate removal (or deboilerplating) is accomplished by applying some form of (usually) supervised machine learning. I.e., an algorithm has to be chosen which is capable of learning a binary decision (a block of text from the document is/is not boilerplate) from a set of training data annotated by humans, and the decision (as learned) should be generalizable from the training set to any new unknown data of the same type. The algorithm learns the decision based on a set of features which have to be extracted for each of the blocks in the training set and for the previously unknown data in a production run." (Schäfer & Bildhauer, 2013, p. 50)

A few leading alternatives are listed in the following paragraphs.

NCleaner [14] (Evert, 2008) uses character n-grams language models. The model is constructed using machine learning on reference data. It is considered to be suitable for many languages but only with sufficient training data since the default model is far from being perfect (Lejeune, 2013).

WCCleaner (Baisa, 2009) is an example of the numerous web page cleaning projects[15] showing that the quality still interests the research community. This system is representative of an alternative way to conceive the problem: the website-level is exploited in order to detect regularities among a particular website.

Readability is available through an Application Programming Interface (API)[16]. The notion of candidate is a key concept in this approach. In a way similar to Boilerpipe, the system uses a list of characteristics to establish whether a particular segment is informative or not, if it is a candidate. Readability adds the idea that good candidates will be found in similar positions in the documents.

[13] http://www.mbayer.de/html2text/

[14] http://webascorpus.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_10_Software

[15] http://is.muni.cz/th/139654/fi_m/

[16] http://lab.arc90.com/

Boilerpipe (Kohlschütter, Fankhauser, & Nejdl, 2010) is available freely on the web[17]. Within the Natural Language Processing community, this is often considered to be the best freely available system (Lejeune, 2013). It uses a segment-based approach and a list of criteria in order to detect if a particular segment is informative or not. Although it does not use the structure properties for cleaning the source code, its output keeps enough layout properties for downstream applications.

Conclusion All in all, statistical patterns are in the lead in order to be able to perform the whole operation, i.e. HTML-stripping and removal of linguistically irrelevant material- automatically, and on a large scale. It is hard to assess the precise impact of boilerplate removal, although it is obvious that it has a real influence on research material, and as such can lead to a distortion of results.

3.3.2.4 Text inclusion into the final corpus

I call text inclusion the decision process which comes after boilerplate removal in web corpus construction, and which is similar to its predecessor in the sense that it is binary. Text parts or whole texts can be either discarded or included in the final corpus. This step is not strictly mandatory, nonetheless it is necessary for most linguistic tasks since it may enhance corpus quality significantly. However, there is no common methodology concerning inclusion of text across general-purpose web corpus projects.

For the Leipzig Corpora Collection project, Biemann, Heyer, Quasthoff, and Richter (2007) perform the task on sentence level, i.e. the documents are segmented into sentences and then the decision takes place for every single sentence. More specifically, a language identifier, which is based on the most frequent 5000 words for each of the languages known to the system, is used in combination with pattern-based methods designed to remove "most of the non-sentences" (Biemann et al., 2007, p. 5). This is possible because the sentences are not recombined afterward to form texts, on the contrary they are given a random number in order to "scramble" the corpus (Biemann et al., 2007, p. 5). In fact, corpus access is sentence-based due to copyright restrictions.

Regarding the WaCky project, Baroni et al. (2009) use pre-compiled lists of frequent words (more precisely function words[18]) as well as fixed thresholds in order to determine if a given text should be included in the corpus. It is a measure relying on type/token characteristics with respect to an expected standard compiled from standard text genres:

> "Documents not meeting certain minimal parameters – 10 types and 30 tokens per page, with function words accounting for at least a quarter of all words – are discarded" (Baroni et al., 2009)

> "The cleaned documents are filtered based on lists of function words [...]
> The filter also works as a simple and effective language identifier." (Baroni et al., 2009, p. 215)

In fact, language identification is a crucial task, which can be tackled efficiently and with an acceptable accuracy by simple indicators, as Baroni et al. (2009) state. However, the difficulty

[17] http://code.google.com/p/boilerpipe/
[18] "Lists of function words (124 items for German, 411 for Italian and 151 for English)"

of the task raises sharply when dealing with multilingual documents or texts designed to fool simple detectors (see chapter 2 for precise examples).

As there are, to my best knowledge, no comparative studies of possible cues so far for web document inclusion or rejection, I tackle this topic below. On one hand from a theoretical, formal point of view, with a conceptual frame for text categorization (see p. 186), and on the other hand in an applied perspective. To this end, decision criteria are analyzed in a controlled experiment (see p. 187).

3.3.2.5 Duplicate and near-duplicate removal

Introduction As one performs larger crawls, it is common to find duplicate contents. In fact, there are many duplicate documents on the Web. This phenomenon can first be explained by the fact that there are more URLs as there are documents:

> "Some duplication stems from the fact that many web sites allow multiple URLs to refer to the same content, or content that is identical modulo ever-changing elements such as rotating banner ads, evolving comments by readers, and timestamps." (Olston & Najork, 2010, p. 225)

Second, copying or mirroring content for legitimate or less legitimate reasons is common practice on the Web:

> "Another source of duplication is mirroring: Providing all or parts of the same web site on different hosts." (Olston & Najork, 2010, p. 226)

Third, due to the rise of Web 2.0 and the injection of content from multiple sources, as evoked in chapter 1, the amount of near-duplicates is also on the rise. Near-duplicates are documents which share, for example, 90% of their content.

Methodological consequences There are several methods in order to deal with this phenomenon. Most of the researchers in the field consider addressing the issue of duplicates as a preprocessing step, and thus perform such an operation before the corpus is released. One should note that for some usages duplicates do not alter the corpus quality. Linguists who are looking for examples may be intrigued by strictly identical sentences coming from several websites, but they are unlikely to reject the corpus just because of that. On the opposite side, the presence of many duplicates can introduce a dangerous bias for statistical approaches, since it means that the words and collocations they contain will be overrepresented.

Method 1: discard all duplicates The proponents of the WaC-corpora (Baroni et al., 2009) have decided to discard all duplicate content, which is justified by technical reasons (it is easier to implement) as well as by a certain defiance to duplicates, which may indicate that content quality is not so high as in "unique" documents:

> "We also spot and remove all documents that have perfect duplicates in the collection (i.e., we do not keep any instance from a set of identical documents)." (Baroni et al., 2009, p. 214)

Method 2: duplicate detection on sentence-level The proponents of the Leipzig Corpora Collection (Goldhahn et al., 2012) have released their corpus as a series of sentences. Thus, they do not need to perform any duplicate removal on document-level, they only do it on sentence-level. Such a practice has the advantage of solving the issue of near-duplicates as well, since most duplications are difficult to detect on document-level, as only a small fraction of the document varies, but easy to detect on sentence-level, because it is trivial to tell if sentences are identical or not. The question of near-duplicate sentences remains unanswered by this approach. To our best knowledge, it is unclear to what extent they are a problem.

Method 3: perfect duplicates removal The method I have mostly used in my work on corpora is a perfect-duplicate removal, a rather trivial operation performed on document-level using efficient hashes. While it does not address the issue of near-duplicates, it has the advantage of functioning without errors: since only perfect matches will be removed, no document will be deleted by mistake. Besides, in restricted retrieval, website structure and corpus properties can be determined in advance, thus making decisions on a case-by-case basis possible.

State-of-the-art, large-scale, efficient, but computationally intensive methods Groundbreaking work on near-duplicate detection has been done by Broder, Glassman, Manasse, and Zweig (1997), originally with the opposite goal in mind, i.e. clustering web documents according to their proximity. The so-called shingling method involves computing similarities across a whole document collection, which on web scale can be a computationally intensive operation.

Several approaches have been developed in order to address the exponentially growing difficulty of the task as web document collections become larger. There are for example RAM-intensive, IO-intensive, or scalable methods. A comprehensive evaluation has been performed by Henzinger (2006). For the use of shingling in web corpus construction, see (Schäfer & Bildhauer, 2013).

The main potential drawback of near-duplicate document removal is that there is no certainty regarding the validity of the operation. A similarity coefficient is computed for document pairs, and the corpus builders are left to decide where to put the threshold in order to cut the collection into two parts, the one which will make it into the final corpus and the one, supposedly made up of near-duplicates, which will be disposed of.

A difficult problem arises from the nature of dynamic web pages, where the content provided by external sources may change anytime. That makes it difficult to find an efficient threshold, since a page which is 98% similar to another one on the same website may actually be exactly the same, the difference of 2% being caused by varying ads or partner links.

Summary Duplicates are very frequent on the Web, making them a relevant issue in web corpus construction. However, the problem of duplicate detection is complex enough to deserve a dedicated study, which is why it is not mentioned further in this document. While perfect duplicates can be found quite easily and without possibility of error, near duplicates cannot be detected with perfect accuracy. Their existence involves testing their potential frequency and using a similarity measure in order to exclude documents from the corpus, which is usually done by comparing series of tokens, or n-grams, and establishing a threshold.

3.3.3 Impact on research results

3.3.3.1 Quantitative impact

A measure of the impact As an example of the tremendous impact of preprocessing, the breakdown performed on the DECOW2012 corpus in several stages (Schäfer & Bildhauer, 2013) shows how much web text is discarded in a conservative approach. In fact, due to the necessity of delivering a high-quality corpus to theoretical linguists, as much as 94% of the web documents which have been downloaded and stored are eliminated during the procedure.

Algorithm removes	No. of documents	Percentage
Very short pages	93,604,922	71.67%
Non-text documents	16,882,377	12.93%
Perfect duplicates	3,179,884	2.43%
Near-duplicates	9,175,335	7.03%
Total	122,842,518	94.06%

Table 3.1: Amount of documents lost in the cleanup steps of DECOW2012, a 9 billion token corpus of German from a crawl of the `.de` domain, according to Schäfer & Bildhauer (2013, p. 19)

Table 3.1 shows that a lot of computational effort is lost during preprocessing, which means that the infrastructure costs are really higher than they could be.

Raising awareness about consequences on the content Despite the cleaning steps, web corpora still aggregate a certain amount of undesirable material. In the eyes of traditional corpus linguists, this may even be their worst downside.

> "[Web corpora] still contain significant amounts of noisy data, such as spam Web pages, redundant auto-generated content from content management systems, misspellings, etc. Compared to users of traditional corpora, users of Web corpora must therefore be more aware of the steps which were taken in the construction of the corpus, such that they are aware of potential distortions of their results." (Schäfer & Bildhauer, 2013, p. 6)

Variations in content quality can have consequences even for the most enthusiastic user base. An advantage which computational linguists see in web data resides in the mere quantity of available occurrences, in the sense that more unclean data seems to be better than less clean data, which could per se enable statistical methods such as language models to obtain better results.

However, this maxim should not be adopted systematically, and a number of statistical measurements should be taken with caution and not compared with more traditional corpora without taking potential differences into account:

> "As an example, [...] the word type count for Web corpora is usually much too high to be plausible due to a large amount of noisy material, such that naively drawn statistical conclusions might be invalid." (Schäfer & Bildhauer, 2013, p. 6)

The more complicated the processes, the less directly interpretable the results The issue of data "clean up" is mentioned by S. Abney (1996) when talking about statistical inquiries:

> "There is always the danger that the simple principles we arrive at are artifacts of our data selection and data adjustment" (S. Abney, 1996, p. 11)

Even from a statistical point of view, access to the corpus is far from being immediate, because of the blend of research goals and processing issues described above. This blend is particularly difficult to see through regarding web corpora; a small shift in research practices can hinder a proper interpretation and assessment of results, even if the raw text base is the same:

> "The value of statistical measurements strongly depends on their reproducibility and comparability. Even small changes in used definitions or working steps can lead to uncomparable and unappraisable results. This especially holds for Web corpora with a large set of pre-processing steps. Here, a well-defined and language independent pre-processing is indispensable for language comparison based on measured values. Conversely, irregularities found in such measurements are often a result of poor pre-processing and therefore such measurements can help to improve corpus quality." (Eckart et al., 2012, p. 2321)

3.3.3.2 Evaluation of boilerplate removal

Usage: corpus construction or readability enhancement for mobile devices (Kohlschütter et al., 2010). Even on mobile devices it is important to preserve the layout as it helps the reader to process the text with an acceptable cognitive cost. Therefore, if a corpus is used for understanding/simulating how humans understand texts, these marks have to be kept.

The CLEANEVAL competition (Baroni, Chantree, Kilgarriff, & Sharoff, 2008) has been an attempt to evaluate several boilerplate removal tools, with major shortcomings. Most importantly, the evaluation metrics favored destructive annotation, since potential loss of original markup was not well evaluated. However, HTML tags convey information helping the reader to get information, so that they ought to be taken into consideration for most applications (Lejeune, 2013).

The main issue resides in the distinction between informative and less (or non-)informative content. In the case of press articles for instance, information is not only conveyed by bare words, since the ability to extract metadata such as title and date is at least as important. Other end-users may also consider paragraph boundaries as highly relevant information.

Most of the time, tools are evaluated on English text, which raises an issue concerning other frequent languages on the Internet such as Russian or Chinese, as well as concerning multilingual corpora.

In this context, the metrics which are chosen also leave matter for discussion. Metrics on word-level are not appropriate for languages like Chinese, where a character n-gram evaluation would be better (Lejeune, 2013).

> "From the point of view of our target user, boilerplate identification is critical, since too much boilerplate will invalidate statistics collected from the corpus and impair attempts to analyze the text by looking at KWiC concordances. Boilerplate stripping is a challenging task, since, unlike HTML and javascript, boilerplate is

natural language text and it is not cued by special mark-up." (Baroni et al., 2009, p. 215)

3.3.3.3 Qualitative impact

Beyond mere data cleaning, a corpus generally aims at being both an authentic and representative sample of language, as proponents of corpus linguistics such as Firth, Halliday or Sinclair "share the belief that each single act of communication shows the language system in operation" (Tognini-Bonelli, 2001).

This issue is all the more important in text linguistics, a case for which the web texts should not be truncated, because variations in the context can go as far as to invalidate interpretation:

> "One fairly obvious feature of a text is that it is not the same all the way through. In barest outline it has a beginning, middle and end, but it is likely to have a much more elaborate structure than that, and each aspect of its internal structure leads to different phraseology, different vocabulary and different structures." (Sinclair, 2008, p. 25)

Then, how should one proceed with web corpora? In fact, cleaning drastically impacts the final collection, so that authenticity can be questioned, which undermines the corpus reasoning process. A proper evaluation process of boilerplate removal, as shown below, and text quality, as shown in the next section, can be crucial.

Analyzing a corpus' quality should take into account the potential corpus users, since there are different understandings of corpus quality, corresponding to diverging requirements among disciplines.

> "There are diverse notions of 'corpus quality' (and, consequently, 'noise'), which depend on the intended use of the corpus. In empirically oriented theoretical linguistics, carefully selected sampling procedures and non-destructive cleaning is important, while for many tasks in computational linguistics and language technology, aggressive cleaning is fundamental to achieve good results." (Biemann et al., 2013, p. 23)

In order to exemplify corpus quality concerns, two examples are given below, with a linguist as potential corpus user in mind.

3.3.3.4 Practical examples from the point of view of a linguist

In the general-purpose approach of corpora from the web, insufficient corpus size is not a problem anymore, the corollary being that text quality becomes one.

> "Crawled raw data for web corpus construction contains a lot of documents which are technically in the target language, but which fail as a text. Documents just containing tag clouds, lists of names or products, etc., need to be removed or at least marked as suspicious. Defining the criteria by which the decision to remove a document is made, however, is quite difficult. For instance, many documents contain a mix of good and bad segments and thus represent borderline cases. The decision to systematically remove documents is thus a design decision with major

consequences for the composition of the corpus and with potential negative side effects on the distribution of linguistic features." (Schäfer et al., 2013, p. 7)

Text selection is a consequence of extreme text and markup variety and also of potential preprocessing pitfalls. Thus, it integrates into a global processing toolchain which goes from the raw HTML document to the annotated, accessible corpus document. It is one of the steps which are performed at some point in each and every web corpus project, but whose impact is often underrated by end users and sometimes corpus designers themselves.

Example 1: tags clouds, lists, and/or search engine optimization techniques The following paragraphs are taken from test data[19] obtained after preprocessing.

A high quality, independent, contract caterer for business, industry, food production and distribution sites. We deliver a hassle-free service, with strict controls ensuring all legal requirements are exceeded. Connect Catering

food service, contract catering, canteen, school catering, independent caterers, industrial catering, university catering, catering contractors, staff canteen, staff restaurant, college catering, connect catering, connect catering ltd, food and service, on site catering, professional catering management, school dinners, schools catering, site catering, university food, works canteen

Example 2: classified ads The following paragraphs are also taken from test data[20] obtained after preprocessing. The text lacks any coherence, because the car ads it is taken from were only partially extracted from a website whose layout is disorienting to say the least.

renault espace 2.0 t privilege 5dr 2004 priviledge 5 door estate, grey, petrol, manual, rear wiper, immobiliser, solid paint, passenger airbag, trip computer, 1 previous owner(s), 2...

scroll over the thumbnails to enlarge model year: 2009 mileage: 31,500 miles transmission: manual engine size (in ccm): 1,995 power: 150 bhp fuel: diesel interested? call us 01209 821133 or 01209 821133 or email us call or visit a friendly member of our sales team who will be more than happy to help. dales motor group wheal rose scorrier redruth cornwall tr16 5bx find out where we are

The classification of the examples above is not as clear as it seems, as discussions regarding these examples showed me (more details in the following section).

[19] Extracted from test data analyzed in (Schäfer et al., 2013)
http://www.cylex-review.co.uk/tags/hygiene,health/?sort=rating_desc

[20] Extracted from test data analyzed in (Schäfer et al., 2013)
http://www.carocean.co.uk/for-sale-Renault+20-6.html

3.4 Conclusive remarks and open questions

3.4.1 Remarks on current research practice

The apparent simplicity of data collection Behind the apparent simplicity of data collection on the Web, there are many mechanisms and biases which can influence the course of a crawl, since a crawl generally means the exploration and retrieval of a tiny sample of the WWW.

> "In a sense, data collection is the simplest step in Web corpus construction [...] However, the pages which are crawled are a sample from a population (the Web documents), and the sampling procedure pre-determines to a large extent the nature of the final corpus." (Schäfer & Bildhauer, 2013, p. 35)

Crawler traps and other deception mechanisms targeted at machines are other potential downsides, which make the way a crawler learns and finds its way through the Web really differs from a general surfer's experience.

Moreover, the issue of corpus representativeness, which for part of the corpus tradition is a desideratum, changes dramatically in its nature with respect to web representativeness. The latter is even harder to define and to analyze than the former, it is a challenge that still remains to be addressed, with projects currently being funded on web corpus sampling and classification.[21]

Big data opportunism In a way, the enthusiasm about huge datasets and the belief that probabilistic approaches can mitigate potential design flaws and inconstant text quality can become an obstacle to corpus approaches on the linguistic side.

There seems to be a "quick and dirty" work tradition concerning web corpora, coming especially from language modeling and machine translation specialists. N-gram-based approaches are no panacea, since a proper cleaning and segmentation remains crucial.

This trend is focused on application and benchmarks such as the translation evaluation frameworks. The ability to gather as much parallel text as possible to run through existing toolchains and finally score through evaluation frameworks supersedes a proper quality evaluation phase.

There is a widely-shared belief among computational linguists working on web corpora that researchers could take advantage of existing linguistic processing and annotation infrastructure.

Web data and web corpus scientists are needed Due to commercial interests, the exact process of web crawling is not well-documented, there is no precise manual one could take inspiration from, and before 2009-2010, there did not seem to be any synthetic overview of what web crawling is and how it is done. In this context, notions of Web science can be very helpful in order to understand what happens during a crawl, so that the final result is not left to chance.

All in all, there is a real need for skilled data scientists, combining the skills of computer scientists and librarians (The Royal Society Science Policy Center, 2012, p. 64).

[21] http://gepris.dfg.de/gepris/projekt/261902821

> "Given the breadth of the Web and its inherently multi-user (social) nature, its science is necessarily interdisciplinary, involving at least mathematics, CS [computer science], artificial intelligence, sociology, psychology, biology, and economics." (Hendler et al., 2008, p. 64)

Web for corpus and Web 2.0: a post-BootCaT world? Web corpora are prone to diverse biases and based on a constantly changing resource. These changes are combined with an evolving web document structure and a slow but irresistible shift from "web as corpus" to "web for corpus", due to the increasing number of web pages and the necessity to extract what is after all a tiny subset of what the web as we know it is supposed to be.

All these changes are part of what I call the post-BootCaT world in web corpus construction (Barbaresi, 2013a).[22]

There are not only fast URL changes and ubiquitous redirections. Content injected from multiple sources is a growing issue for recent general-purpose web corpora which can be linked to the Web 2.0 paradigm (see p. 116), because it may be that the probability of running into lower text quality, for instance because of machine-generated content or mixed-language documents, as described above, increases with the number of different sources. Additionally, external sources changing at a fast pace make a filtering or blacklisting of domain names more difficult to implement.

As the complexity of a document with respect to its source(s) rises, so does the difficulty to establish a quotable source for linguists depending on a more traditional reviewing process of linguistic proof and who need reliable, clearly established sources. It also makes the decision to exclude potentially noisy sources more delicate.

3.4.2 Existing problems and solutions

Recent advances Recent advances in general-purpose corpora, for instance in the work of Suchomel and Pomikálek (2012) or Schäfer and Bildhauer (2012), include resource-efficient processing tools, steps towards encoding of available metadata, and overall cleaner corpora through better selection and verification procedures. The crawling infrastructure, corpus processing tools, and corpus search engines make it possible to create and master web corpora on a scale of 10 billion tokens or more, for languages with a large, worldwide speaker community such as French or Spanish, but also for other cases such as Swedish or Czech (Jakubíček et al., 2013).

On the side of specialized corpora, work has been done to help with the normalization and annotation of text types such as Internet-based communication (Beißwenger et al., 2013), for which there are neither annotating schemes nor processing practices. Still, targeting small communities or particular text types, extracting the texts, and processing the documents remains a challenge.

Existing problems Search engines have not been taken as a source simply because they were convenient. They actually yield good results in terms of linguistic quality. The main advantage was to outsource operations such as web crawling and website quality filtering, which are

[22] Note that the proponents of the BootCaT method seem to acknowledge this evolution, see for example Marco Baroni's talk at the BootCaTters of the world unite (BOTWU) workshop (2013): "My love affair with the Web... and why it's over!"

considered to be too costly or too complicated to deal with while the main purpose is actually to build a corpus.

Nonetheless, the quality of the links may not live up to expectations. First, purely URL-based approaches favor speed, sacrificing precision. Language identification tasks are a good example of this phenomenon (Baykan et al., 2008). Second, machine-translated content is a major issue, as is text quality in general, especially when it comes to web texts (Arase & Zhou, 2013). Third, mixed-language documents slow down text gathering processes (King & Abney, 2013). Fourth, link diversity is also a problem, which in my opinion has not gotten the attention it deserves. Last, the resource is constantly moving. Regular exploration and re-analysis could be the way to go to ensure the durability of the resource.

The inefficiency of crawling The crawling process in itself cannot be completely inefficient: since even content prediction using URLs cannot be expected to be accurate, the links have to be visited in order to retrieve the content of a page, if only to discover that there is no or little text content. Once the web documents are stored, up to 94% of the web documents which have been downloaded and stored are eliminated during preprocessing (Schäfer & Bildhauer, 2013).

Thus, crawling and preprocessing are resource-intensive, so that the fact that much computation time could be saved is highly relevant, in particular when processing power is short. Thus, the question is whether crawling (in)efficiency is unalterable or if it can be improved. As it is not possible to start a web crawl from scratch, the question concerns both the sources and the course of the crawl, and can be roughly formulated as such: where may one find web pages which are bound to be interesting for corpus linguists and which in turn contain many links to other interesting web pages?

Preprocessing I used the word preprocessing, and not post-processing as sometimes found in the literature, in order to highlight the fact that it is this operation that makes corpora exploitable and available in a linguistic sense. It is more than just a cleaning operation, as it involves selecting the texts and balancing a corpus between opportunistic inclusion and strict selectivity, thus affecting its general profile and quality.

One may say that preprocessing suffers from a lack of attention since there is no external evaluation procedure as to how useful given web texts are for a corpus, and the last evaluation campaign regarding boilerplate removal dates back from 2008 and leaves much to be desired.

3.4.3 (Open) questions

Answerable questions The following questions are answered at least partially in the following:

- How can the linguistic relevance of web data be assessed?
  Under what conditions is a text worthy to become part of a corpus?

- Is it possible for public research infrastructures to gather large quantities of text from the Web
  In fact, Tanguy (2013) states that publicly funded research centers cannot compete with commercial search engines, mostly because of infrastructure costs. The answer to that problem may be to look for more efficient ways to build web corpora. In the following, solutions are introduced (see chapter 4).

  - Are there ways to cope for the search engine APIs which are being closed when looking for URL crawling seeds?
  - What are possible ways towards more efficient crawls?
  - How can web corpus construction and its issues be brought closer to linguists and not left to “crawl engineers"?

- Operationalize document classification
  Towards the reproducibility of decisions

Open questions The following questions arise from the state of the art presented in the general introduction as well as in this part. They are known to be of interest, but fall beyond the scope of this work.

- How many text genres can be identified on the web?
  What are the best machine learning techniques to deal with these genres?
- What is the best way to find promising and evenly distributed URL crawling seeds?
  What is the best crawling strategy?
- What is the most adequate solution to the debate about corpus balance?
- What is the best boilerplate removal method? How can it be evaluated on a wide range of criteria and texts?

Chapter 4

Web corpus sources, qualification, and exploitation

"Gardens are not made by singing 'Oh, how beautiful!' and sitting in the shade."

— Rudyard Kipling, *Complete Verse*

4.1 Introduction: qualification steps and challenges

4.1.1 Hypotheses

4.1.1.1 From pre-qualified URLs to web corpora

Many problems of web corpus construction have been solved neither conceptually nor technically, for example the issue of web genres as well as the question whether balance is relevant to web corpora. Tanguy (2013) claims that an interesting research program would be to work towards automatic characterization processes that do not classify web pages but rather yield useful information for further linguistic exploitation. As such, the resulting data could then be used by more advanced tools.[1]

In fact, it could be useful to pre-qualify web documents, i.e. work on lists of URLs then used as sources for a crawl, in order to spot "licit" contexts (Tanguy, 2013) for use by corpus linguists, and thus make web corpus building easier. The whole process could be divided into the following main steps:

1. Fit to the peculiarities of "web texts", most notably by finding appropriate text descriptors.
   In fact, the heterogeneity of the raw material makes effortless data mining impossible. It is necessary to perform some data wrangling and to find salient surface cues (such as formal and sentence-based descriptors) in order to manipulate this material:

   > "While in some domains, simple 'data mining' is conceivable, in the case of text corpora (and, possibly, for other 'heterogeneous' databases), prior redescription is a necessity." (Wallis & Nelson, 2001, p. 313)

2. Prequalification step:

   - Filter URLs and web documents using the resulting relevant characteristics that enable the expression of heuristics and statistical processes.
   - Annotate the resources so that they can fit various users' needs.

3. Qualification step: using metadata added on purpose, qualify downloaded web documents, i.e. determine if they seem suitable for inclusion in a corpus.

The operation of qualification may seem similar to a linguistic characterization of the texts. However, due to the extreme diversity of the documents taken into consideration (see definition and examples), the qualification brings statistical significance as well as features related to the operationalization into focus. Robustness for example is paramount. As Bronckart, Bain, Schneuwly, Davaud, and Pasquier (1985) explain, these kinds of results do not yield conclusions at a linguistic level per se (or indirectly at best).[2]

[1] "Une des pistes les plus intéressantes concernerait à mon avis la mise en place de procédures automatiques de caractérisation à la volée, ne visant pas à la catégorisation en genres, mais permettant par contre de donner des informations utiles sur une page Web pour une exploitation linguistique (par exemple l'identification de contextes considérés comme licites). De telles procédures couvriraient des besoins en googleologie, et seraient insérables dans des approches plus lourdement outillées." (Tanguy, 2013, p. 29)

[2] "On ne peut généralement pas inférer de la significativité statistique d'une différence sa pertinence sur le plan linguistique ou plus généralement communicatif." (Bronckart et al., 1985, p. 72)

Nonetheless, this identification should be performed without losing sight of a typological perspective. As Loiseau (2008) explains, automatic classification is not a goal by itself, it is paramount to take a stand on textual typology in order to come to a proper description.[3]

4.1.1.2 Usefulness of the results of readability studies for filtering and qualification tasks

I formulate the following hypotheses with respect to the notions introduced in chapter 2:

- It is possible to develop a filter based on URLs, HTML characteristics and text-based statistics in order to discriminate between incoherent web-specific document types and linguistically relevant ones with a good precision.
- Research on text readability has yielded a series of metrics and analytical tools which can be generalized for text analysis purposes. These results can be used to qualify texts gathered on the web which are annotated on this basis. In fact, indicators can be aggregated to build multi-dimensional criteria that enable a proper classification of the texts and/or the construction of subcorpora.
- The precision as well as the recall of the prequalification and qualification can be evaluated using specially designed samples (respectively URL and text samples) as well as with a series of heuristics applied to the whole corpus, such as, for instance, existing anti-spam tools, n-gram dispersion or language model perplexities.

4.1.1.3 Insights on collected corpora

Concerning the content exploitation as well as exploration, the following hypotheses can be formulated:

- It is possible to develop semi-automatic procedures to help with quality assessment.
- Corpora gathered on the Web can be compared to existing reference corpora as well as to one another, for quality assessment as well as for typological purposes.
- It is possible to give access to corpus texts via a visualization interface.

4.1.2 Concrete issues

4.1.2.1 Qualification of URLs (prequalification)

Problems to address The methodology used so far to gather URLs relies heavily on search engine APIs. However, many APIs are no longer freely available. Moreover, the question whether this is the best way to collect a high number of URLs reflecting language use on the web remains open, as the so-called BootCaT method is prone to several serious biases (see above). Link gathering on other sources like social networks is a way to solve this URL shortage problem and to complement the biases of search engine algorithms and optimization by adding user-based information to the crawl seeds.

[3] "Les nouveaux moyens de description de la textualité doivent donc sans doute être articulés à un programme de typologie textuelle, au-delà des perspectives de classification automatique." (Loiseau, 2008)

If one considers all potential URLs in a breadth-first search manner, a large part is redundant, misleading, or simply does not lead to a kind of content that one would consider integrating in any kind of text corpus.

Therefore, the main goal seems to be a proper calibration of the filter, which should only remove URLs that are obviously not reliable when it comes to creating a text corpus, leaving the rest of the work to a content filter.

In terms of precision and recall, the latter is preferable, as it is more important to keep as many interesting documents as possible, because the ones that lack relevance can always be filtered out later.

Method The prequalification of URLs has recently become a research topic by itself, all the more since big data became a field of interest. Due to the quantity of available web pages and the costs of processing large amounts of data, it has become an Information Retrieval task to try to classify web pages merely by taking their URLs into account and without fetching the documents they link to. Several heuristics such as trigram-based methods have proven to be efficient as a first pass, be it topic (Baykan, Henzinger, Marian, & Weber, 2009) and genre guessing (Abramson & Aha, 2012) or language identification (Baykan et al., 2008).

URL classification has also been used to find parallel texts for example, leaving a lot of questions unanswered as to the text quality of the “dirt cheap" corpus gathered this way (Smith et al., 2013), showing that it is not a trivial task as it impacts all downstream applications.

The work mentioned above paves the way for a first-pass filter enabling selection of possible candidates for a web corpus before actually downloading anything. Spam and advertisement are a major issue, but also simple URLs that lead to image or video files or web pages that do not mainly consist of text, for example photoblogs.

4.1.2.2 Qualification of web texts

Problems to address There are obviously content and text types on the Internet that do not belong to a linguistically relevant text corpus (see examples below).

It is not so easy to filter them out because the inter-annotator agreements are remarkably low (see p. 189). It seems that the extension of the notion of web corpus varies greatly according to the possible end users.

Due to the wide variety of web texts it is necessary to find robust definitions and to build or use robust tools in order to enable classification and quality assessment in a large range of different languages, text genres, and web page types.

Exploitation and visualization Special interfaces are needed in order to provide easier access to the actual content of web corpora, both restricted and general-purpose ones. In fact, due to the size of the corpora, it is not usually possible to read or even skim through a significant part of their content. The main access available even on bare, unannotated texts is to examine random samples of the corpus.

Quality assessment and general corpus analysis could be easier with other ways to look at corpora, for instance from the particular angle of a precise tool, or using a visualization which maps either a particular characteristic as it is present or absent throughout the corpus, or a general summary of corpus content.

4.2 Prequalification of web documents

4.2.1 Prequalification: Finding viable seed URLs for web corpora

4.2.1.1 The FLUX toolchain: Filtering and Language-Identification for URL Crawling Seeds

Description The FLUX-toolchain is a light scout designed to be faster and use less resources than a full-fledged crawler. Its purpose is to tackle the makeup of viable URL lists which are to be used as the start of a crawl. These URLs can be seen as more “promising" than random URLs because they are checked against a range of characteristics. Additionally, the links they contain are also analyzed.

As such, it has to handle several kinds of problems in order to move obstacles out of the way. First, the very validity of the URLs have to be checked, i.e. whether they actually lead to web documents. Redirection checks are performed for example. Then, operations on the domain name take place, for example the search for spam, and also a control of the distribution of URLs, so that a certain diversity can be enforced.

The actual documents are retrieved in order to ensure that they are suitable. Several kinds of factors are taken into account: technical ones, e.g. the actual response of the web server, superficial ones, such as the length of the text of the web page, and linguistic ones, such as the main language the web page is written in.

All in all, the light scout is expected to yield a URL directory which yields a reasonable image of the content a crawler is bound to run into. Based on the metadata gathered for each URL in the directory, seed URLs for a crawl can be extracted. The rationale for such a preliminary step is based on three main hypotheses:

1. The BootCaT method (see above) can be complemented or replaced if necessary.
2. It is possible to perform such a step using relatively simple and cost-effective methods.
3. It can be shown that crawls starting with prepared seed URLs lead to more effective results.

First, the toolchain is presented. Second, several cases for alternative URL sources are introduced, together with an evaluation of results. Finally, main existing URL sources are compared using the FLUX-method. In the next section, impact on crawling processes is studied.

Steps of the toolchain The following sketch describes how the results below are obtained:

1. URL harvesting: queries or archive/dump traversal, filtering of obvious spam and non-text documents.
2. Operations on the URL queue: redirection checks, sampling by domain name.
3. Download of the web documents and analysis: collection of host- and markup-based data, HTML code stripping, document validity check, language identification.

Figure 4.1 offers a graphical summary of the components of the processing chain.

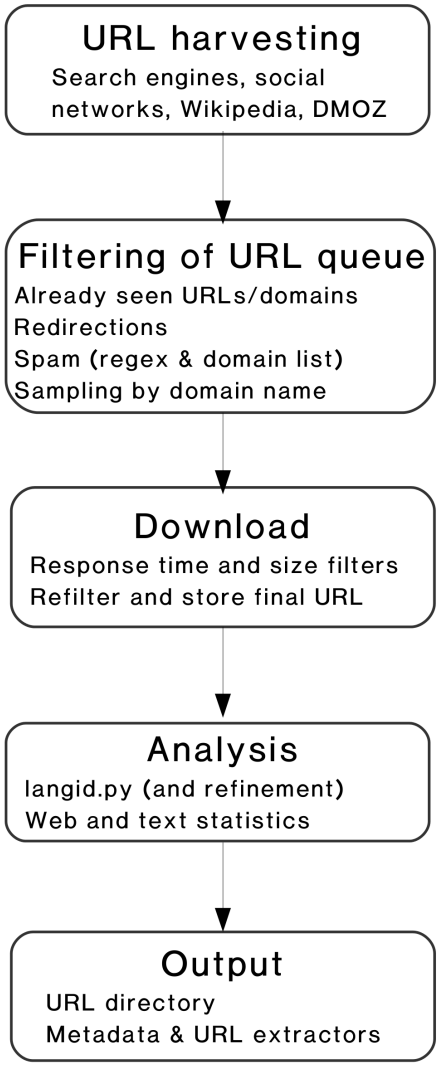


Figure 4.1: Main components of the FLUX toolchain

Details of URL filtering As a page is downloaded, links are filtered on the fly using the series of heuristics described below. If several URLs point to the same domain name, the group is reduced to a randomly chosen URL. This sampling step reduces both the size of the list and the potential impact of overrepresented domain names in final results.

Links pointing to media documents have been excluded from the studies, as the final purpose of these studies is to allow for the construction of a text corpus. The URL checker removes non-http protocols, images, PDFs, audio and video files, ad banners, feeds and unwanted hostnames like *twitter.com*, *google.com*, *youtube.com* or *flickr.com*. Additionally, a proper spam filtering is performed on the whole URL (using basic regular expressions) as well as at domain name level using a list of blacklisted domains comparable to those used by e-mail services to filter spam. As a page is downloaded or a query is executed, links are filtered on-the-fly using a series of heuristics described below, and finally the rest of the links are stored.

There are two other major filtering operations to be aware of. The first concerns the URLs, which are sampled prior to the download. The main goal of this operation is strongly related to my scouting approach. Since I set my tools on an exploration course, this allows for a faster

execution and provides us with a more realistic image of what awaits a potential exhaustive crawler. Because of the sampling approach, the “big picture" cannot easily be distorted by a single website. This also avoids “hammering" a particular server unduly and facilitates compliance with *robots.txt* as well as other ethical rules. The second filter deals with the downloaded content: web pages are discarded if they are too short. Web documents which are more than a few megabytes long are also discarded.

The first step of operations on the URL queue (cf steps above) consists of finding the URLs that lead to a redirect, which is done using a list comprising all the major URL shortening services and adding all intriguingly short URLs, for example URLs which are less than 26 characters in length (value determined empirically). To deal with shortened URLs, one can perform HTTP HEAD requests for each member of the list in order to determine and store the final URL.

The second step is optional, and comprises a sampling that reduces both the size of the list and the probable impact of an overrepresented domain names in the result set. If several URLs contain the same domain name, the group is reduced to a randomly chosen URL. Algorithm 1 describes a possible way to sample the URLs.

Algorithm 1 Sampling of URLs

Require: a sorted URL list where all the lines are unique
  while there are links to examine **do**
    extract the link's hostname
    if hostname $\neq$ last seen hostname **then**
      store a randomly chosen link from the temporary list in the primary list
      clean the temporary list
    else
      store the link in a temporary list
    end if
  end while
Ensure: save the primary list to a file

There are two advantages of stripping the path: sometimes a path and the bare hostname lead to different parts of a website that are written in different languages. Moreover, the hostname is not always included in the list, but may contain information. If there is little or no text, it will not be taken into consideration for further crawling steps.

Due to overlaps of domain names and the amount of spam and advertisement on social networks such an approach is very useful when it comes to analyzing a large list of URLs.

Moreover, a proper spam filtering is performed on the whole URL (using basic regular expressions) as well as at domain name level using a list of blacklisted domains comparable to those used by e-mail services to filter spam.

Details of web document analysis Regarding the web pages, the software fetches them from a list, strips the HTML code, sends raw text to a server instance of *langid.py*, the language identification software described below. It then retrieves the answer, on which it performs a sanity check.

Optional language identification using a spell-checker First, a quick test can be performed in order to guess whether a text is English or not. Indeed, this operation cuts the amount of texts in half and enables to select the documents featuring the desired response, thus directing the analysis in a more fruitful direction.

The library used, `enchant`, allows the use of a variety of spell-checking backends, like aspell, hunspell or ispell, with one or several locales.[4] Basically, this approach can be used with other languages as well, even if they are not used as discriminating factors in this study. We consider this option to be a well-balanced solution between processing speed on the one hand and coverage on the other. Spell checking algorithms benefit from years of optimization in both areas.

This first filter uses a threshold to discriminate between short messages, expressed as a percentage of tokens which do not pass the spell check. The filter also relies on software biases, like Unicode errors, which make it nearly certain that the given input microtext is not English.

Language identification with langid.py I consider the fact that a lot of web pages have characteristics which make it hard for "classical" NLP approaches like web page language identification based on URLs (Baykan et al., 2008) to predict the languages of the links with certainty. That is why mature NLP tools have to be used to qualify the incoming documents and enable a language-based filtering based on actual facts.

A language identification tool is used to classify the web documents and to benchmark the efficiency of the test mentioned above. `langid.py` (Lui & Baldwin, 2011, 2012) is open-source[5], it incorporates a pre-trained model and covers 97 languages, which is ideal for tackling the diversity of the web. Apart from this coverage, the software is versatile. We used it as a web service, which made it a fast solution enabling distant or distributed work.

The server version of `langid.py` was used, the texts were downloaded, all the HTML markup was stripped and the resulting text was discarded if it was less than 1,000 characters long. According to its authors, `langid.py` could be used directly on microtexts. However, this feature was discarded because it did not prove as efficient as the approach used here when it comes to substantial amounts of short messages.

As the software is still under active development, it can encounter difficulties with rare encodings. As a result, the text gets falsely classified, for example as Russian or Chinese. The languages I studied are not affected by these issues. Still, language identification at document level raises a few problems regarding "parasite" languages (Scannell, 2007).

4.2.1.2 The case of low-resource languages

Aim of the study In this subsection I will report the results of my experiments[6] regarding the evaluation of several web corpus construction strategies for low-resource languages (see above for a definition). With these experiments I wish to highlight the challenges linked to the peculiarities described above and find novel ways to access the resources (which in this case are the web texts), such as the social network exploration I also implemented (Barbaresi, 2013b) (see p. 160).

[4] `http://www.abisource.com/projects/enchant/`

[5] `https://github.com/saffsd/langid.py`

[6] This work has been partially funded by an internal grant of the FU Berlin, COW (COrpora from the Web) project at the German Grammar Department.

The main issue I would like to address concerns post-BootCaT web text gathering: What are viable alternative data sources for low-resource languages such as Indonesian? I think that established directories could yield better results than a crawl “into the wild", with advantages such as spam avoidance, diversity of topics and content providers, and better quality of content.

To do so, I have implemented the first exploration step that could eventually lead to full-fledged crawls and linguistic processing and annotation: a light scout enabling the discovery of resources and building of a language-classified URL directory. Besides, my experiments also make it possible to see how far one may go using different types of sources. The whole process gives an insight about the linguistic nature of the afferent resources and about the challenges to address when exploring a given web space.

In the paragraphs below, I will introduce my experimental setting, i.e. the studied languages, data sources and goals. Then I will describe the metrics used to try to evaluate the resources. In section four I will list and discuss the experimental results, and make a conclusion by summing up the challenges I have cast light on.

Languages studied: Indonesian, Malaysian, Danish, and Swedish My research interest originates in a paradox: “Large standard languages – those with numbers of native speakers in the hundreds and tens of millions and having a long tradition of writing – are not necessarily high- or even medium-density languages" (Borin, 2009a).

In order to study this problem I chose two languages with a low “resource to population size ratio" on one side and on the other two languages presumably very different from this perspective. I focused primarily on the Indonesian language which in my opinion is a significant example, as it should not at all fall into the under-resourced languages category: according to census data[7], there are more than 60 million Internet users in Indonesia alone, which leaves a substantial number of users writing or reading primarily in this language, even if one takes into account the multiethnicity of Southeast Asia.

Questions linked to Indonesian arose from previous studies and global web crawls, during which I only found a few websites in Indonesian. I propose that in spite of the potential number of Internet users, the Indonesian Web is not well connected to the Western world, from a technical as well as from a cultural interlinking point of view, so that the chances of finding Indonesian pages during a typical crawl are scarce.

Indonesian (Bahasa Indonesia) and Malaysian (Bahasa Malaysia) are closely related. The pair Indonesian/Malaysian is mentioned by Scannell (2007) as being part of the under-resourced languages but also as a language pair that is difficult to distinguish. Thus, it is important to consider both languages at once because it is sometimes difficult to draw a sharp line between their linguistic variants, all the more so for the language identification tools.

I have performed all studies on Indonesian and some on Malaysian, taking the language pair into account during the interpretation process. In order to have a point of comparison, I have chosen a Scandinavian language pair, Danish and Swedish. When it comes to written texts, these two languages are probably easier to distinguish. In fact, they are medium-resourced languages and not low-resourced languages, which has an impact on production processes and epilinguistic knowledge on the one hand, and on language identification on the other. First, the speakers are supposed to be aware that they are writing in Swedish or Danish, and second, the resources for building tools for these languages are more numerous and more stable.

[7]Population of 237,424,363 of which 25.90% are Internet users. Data from 2011, official Indonesian statistics institute (http://www.bps.go.id).

Data sources In order to perform a comparison I have chosen two main data sources. First of all, the Open Directory Project (DMOZ)[8], where a selection of links is curated according to their language or topic.[9] The language classification is expected to be adequate, but the amount of viable links as well as the content is an open question: What are these URLs worth for language studies and web corpus construction? I have analyzed the directory itself as well as the possible results a crawl using these web sites might obtain.

The free encyclopedia Wikipedia is another spam-resilient data source where the quality of links is expected to be high. It is known that the encyclopedia in a given language edition is a useful resource. The open question resides in the outlinks, as it is hard to get an idea of the global picture due to the number of articles: do the links from a particular edition point to relevant web sites (with respect to the language of the documents they contain)? I have classified these outlinks according to their language to try to find out where a possible crawl could lead.

Processing pipeline In the context of the Indonesian language, I agree with Scannell (2007): it is clearly inefficient to crawl the web very broadly and to only filter by language at the end of a crawl. Thus I have adopted a similar methodology during the crawling process: parallel threads were implemented, the results are merged at the end of each step, and only the documents in the target language are considered for link extraction, before the retrieval of web pages one depth level further.

Web page and corpus size metrics Web page length in characters is used as a discriminating factor. Web pages that are too short, i.e. less than 1,000 characters long after HTML stripping, are discarded in order to avoid documents containing just multimedia (pictures and/or videos) or, for example, microtext collections, as the purpose is to simulate the creation of a general-purpose text corpus.

The page length in characters after stripping is recorded, so that the total number of tokens of a web corpus built on this basis can be estimated. The page length distribution is skewed, with a majority of short web texts and a few incredibly long documents at the end of the spectrum (see above), which is emphasized by the differences between mean and median values used in the results below.

Host sampling is a very important step of the workflow because the number of web pages is drastically reduced, making the whole process feasible and more well-balanced, i.e. less prone to host biases. IP statistics corroborate this hypothesis. Freshness and in- and outlinks are also handy options when dealing with major languages. However, nothing has been filtered on this side, so the web page discovery would not be hindered.

The deduplication operation takes places at document level using a hash function. The IP diversity is partly a relevant indicator in this case, as it can be used to prove that not all domain names lead to the same server. However, it cannot detect the duplication of the same document across many different servers with different IPs, which in turn the basic deduplication is able to reveal.

Language identification The language identification software `langid.py` is used (see p. 154). Since the software is still being developed, there are difficulties with rare encodings. In this

[8] http://www.dmoz.org/

[9] Thanks to Roland Schäfer (FU Berlin) for the script extracting DMOZ URLs.

study, neither Indonesian nor Malaysian were affected by these technicalities.

Language identification at document level raises a few problems regarding “parasite" languages (Scannell, 2007) such as ads in another language (Baker et al., 2004). However, using a language identification system has a few benefits. It enables to find “regular" texts in terms of statistical properties and exclude certain types of irregularities such as encoding or markup problems since web texts are straightened out. This underlying classification is an interesting property.

Results Table 4.1 summarizes the results for the four languages studied and the two different source types (DMOZ and Wikipedia).

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
DMOZ							
Indonesian	2,336	1,088	71.0	5,573	3,922	540,371	81.5
Malay	298	111	59.5	4,571	3,430	36,447	80.3
Danish	36,000	16,789	89.6	2,805	1,652	5,465,464	32.6
Swedish	27,293	11,316	91.1	3,008	1,838	3,877,588	44.8
Wikipedia							
Indonesian	204,784	45,934	9.5	6,055	4,070	3,335,740	46.3
Malay	90,839	21,064	3.5	6,064	3,812	548,222	59.1
Danish	161,514	33,573	28.3	4,286	2,193	5,329,206	38.1
Swedish	320,887	62,773	29.7	4,058	2,257	8,388,239	32.7

Table 4.1: URLs extracted from DMOZ and Wikipedia

Commentary on DMOZ results First of all, it is noteworthy that the dropped URLs ratio is equivalent in both cases, with about 40% of the URLs being retained after processing (and most notably after domain name sampling). This figure reflects the quality of the resource, as the websites it leads to are expected to be diverse. This is where the IP diversity indicator proves to be relevant, since it confirms this hypothesis. It is interesting to see that the Scandinavian web space seems to have more servers in common than the Indonesian one. This is probably due to a market trend concerning web space rental.

As could be expected due to the specialization of the sources, the majority of web pages are in the target language, all the more if the concurrent pair Indonesian–Malay is considered, with about 15% each time in the concurrent language (a complementary information to the results in Table 4.1). Nonetheless, the difficulty of finding documents in Indonesian is highlighted by these results, where the comparison with Danish and Swedish is highly relevant: there are far more URLs to be found, and the corpus size based on DMOZ alone is roughly ten times bigger.

Commentary on Wikipedia results The ratio of retained to analyzed URLs is lower here, but still constant across the languages studied at about 20%. This still indicates that Wikipedia is a preferred source considering the diversity of the domain names the encyclopedia points to.

The proportion of web pages in the target language is a clear case for the scarcity of resources in Indonesian and Malay. English is found in about 70% of the URLs, and it still amounts to about 45% of the URLs for the Scandinavian language pair.

The average web page seems to be a bit longer, and the mere number of links makes a difference, so that the potential corpora based on Wikipedia contain more text. The drop concerning IP diversity may be correlated to the amount of URLs and may converge at about 30%, as there are not so many website hosters after all.

Crawling experiments The crawling experiments summarized in Table 4.2 show that DMOZ and Wikipedia are good starting points to begin a web crawl. In fact, although the web pages are sampled by domain name, a reasonable amount of URLs is to be reached in three or four steps. Among these URLs, a slightly higher proportion of URLs is retained, showing that the domain name diversity of these steps is still growing. Only the IP diversity is dropping, while the page lengths are in line with the expectations based on the respective start URLs.

Source	Depth	URLs		% in target	Length		Tokens (total)	Different IPs (%)
		analyzed	retained		mean	median		
DMOZ	3	32,036	14,893	34.7	6,637	4,330	4,320,137	34.0
Wikipedia	4	95,512	35,897	24.3	6,754	3,772	7,296,482	28.8

Table 4.2: Crawling experiments for Indonesian

The crawl started with Wikipedia really benefits from the language filtering at each step. By contrast, the drop in percentage of URLs in Indonesian in DMOZ is once again significant. Even when staying focused is the priority, web texts written in Indonesian seem relatively hard to find. This fact explains why target-specific strategies may be necessary. To sum up, the figures confirm that web crawling is definitely an option when it comes to gather larger amounts of text, as the number of tokens increases notably.

The results for experiments summarized in table 4.3 show that DMOZ and Wikipedia are much more efficient as a source than newspaper websites, i.e. the type of approach used by Biemann et al. (2007), even if the results given here are excessively “naive" as they consist of a single extensive breadth-first search crawl.

This is particularly clear in the case of Kaskus, i.e. kaskus.co.id, a community website among the most used websites in Indonesia according to alexa.com. The fact is that the servers performing the crawl had a huge negative impact on the crawl, with 52.1% of the final pages being in German, even though the most frequent concurrent language in the other experiments is English. This failure can be explained by two factors: on the one hand the relatively poor results in terms of links outside of the original domain, which force more crawling steps than it would be advisable to do, and on the other hand the role of website monetization, since advertisement and user-targeted content can account for the high proportion of German pages.

Meanwhile, a selection of about a dozen of top Indonesian blogs performed comparably to the DMOZ and Wikipedia, which encourages seeing the blogs as a valuable resource. The fact

that there were more links in the start URLs concerning the blogs also highlights the necessity to have a significant number of high quality sources, even a dozen could suffice. The blogs also seem to link to websites containing more text than in the other experiments. That said it is not necessarily surprising to see much text in the blogosphere. The main drawback is the relative redundancy, with many URLs coming from the same domains and pointing to the same IP.

Source	Depth	URLs		% in target	Length		Tokens (total)	Different IPs (%)
		analyzed	retained		mean	median		
Blogs	4	89,382	17,660	39.4	13,974	5,290	12,275,512	27.0
Newspapers	6	73,642	14,448	13.1	10,185	3,875	2,331,035	39.7
Kaskus	8	60,289	13,304	2.0	10,496	3,848	318,940	40,8

Table 4.3: Naive reference for crawling experiments concerning Indonesian

Discussion The confrontation with a constantly increasing number of URLs to analyze and the necessarily limited resources make website sampling by domain name useful, as it highlights the challenges in Indonesian web text collection.

A common practice known as cloaking clearly hinders the crawls: a substantial fraction of web servers show a different content to crawler engines and to browsers. This Janus-faced behavior tends to alter the language characteristics of the web page in favor of English results, or even of results in the language of the country which the crawler appears to come from.

In fact, I have also tried to take country-specific popular web sites into account by starting crawls from large portals and blogs, basing on data gathered by the traffic analyzer Alexa.[10] However, this leads to a high proportion of websites being in German after only a few crawling steps, which shows that starting with a low number of URLs, however influential they are, delivers highly variable results. Therefore, it should be avoided. URL diversity is not only useful from a linguistic but also from a technical point of view.

Additionally, in order to better explore the web space corresponding to a given target language, it could prove very useful to determine or spoof the server location accordingly, as this could improve both the retrieval speed and the content language.

From the output of this toolchain to a full-fledged web corpus, other fine-grained instruments, as well as further decisions processes (Schäfer et al., 2013) are needed along the way. As a consequence, future work could include a few more linguistically relevant text quality indicators in order to fully bridge the gap between web data, NLP, and corpus linguistics. I am in favor of the idea that corpus building is similar to language documentation as described by (Austin, 2010), since it requires a scientific approach to the environmental factors during information capture, as well as data processing, archiving, and mobilization.

The information I have collected raises the awareness of the proper conditions for information capture. If it was maintained on a regular basis and enriched with more metadata, the URL database I have described could offer a similar approach to data archiving and mobilization. In fact, it could be used as a source for URL crawling seeds in order to retrieve texts based on particular criteria, which could lead to an enhancement of web corpus quality and also to a

[10] http://www.alexa.com/

better suited crawled corpus, according to the hypothesis that linguistically relevant pages are somehow linked to each other.

Conclusion I have evaluated several strategies in order to complement or replace search engine queries to find texts in a given low-resource language. I have shown a possible method for gathering a corpus using two different sources. It leads to a satisfying rate of representation of different hosts, meaning the size of the corpus could increase drastically if one was to remove the sampling process concerning domain names. The scouting approach actually leads to a resource database which can be used to suit particular needs like balanced and/or wide-ranging corpora.

I will extend what Scannell (2007) says about linguistic knowledge by adding that crawling without expert web science knowledge is also "doomed to failure", or more precisely doomed to massive distortions of results, which can and will impact downstream linguistic studies.

4.2.1.3 Exploring microblogging services

Introduction In order to find URL sources other than the search engines used in the BooTCaT approach, social networks and microblogging services seem to be a promising option. Indeed, my hypothesis states that microblogging services are a good alternative for overcoming the limitations of seed URL collections and the biases implied by search engine optimization techniques, PageRank and link classification.

However, the two most prominent websites, Facebook and Twitter, make messages gathering on a large scale a costly and/or hazardous enterprise. Thus, it is necessary to turn to lesser-known websites, which might have two main drawbacks: a sociological bias which is difficult to apprehend, and an even greater instability of web presence and practices.

The present section will describe ways to cope with the difficulties related to social networks as well as results concerning the variety of content which can be gathered on them.[11]

User-based URL gathering The URL gathering process described here uses a user-based language approach.

From a linguistic point of view, it is an interesting process that can give the research community access to actual utterances by a potentially large array of speakers. Such a collection is not new in and of itself, but it was difficult to achieve before the Internet, and more precisely before blogs and microblogs existed. Due to their growing popularity, it is now possible at least theoretically to study language as it is written by a majority of speakers and/or to differentiate between several speaker types.

From a technical point of view however, things are not so simple. Microblogs are a good example of one of the main arguments of this thesis, i.e. the fact that it is necessary to bridge the gap between constraints or expectations on the linguistic side, and practical difficulties or laissez-faire on the side of NLP approaches.

Generally, the amount of spam and advertisement is an obvious practical limit of social networks. For various reasons, content cannot be accessed directly and has to be filtered. Concerning language theory, the user profiles may be a limitation. From a more sociological point of view, the main cause for bias is indeed the technology-prone users who are familiar

[11]This work has been partially funded by an internal grant of the FU Berlin (COW project at the German Grammar Dept.).

with these platforms and produce numerous short messages which in turn over-represent their own interests and hobbies.

Nevertheless, user-related biases also have advantages, most notably the fact that documents that are most likely to be important are being shared, which has benefits when it comes to gathering links in lesser-known languages, below the English-speaking spammer's radar.

Arguments towards languages and website diversity Microblogging platforms may allow for the gathering of higher proportions of URLs leading to lesser-known languages. Following this, I would like to test whether social networks and microblogging services can help focus on them.

In fact, it can be argued that the most engaged social networking nations do not use English as a first communicating language.[12] In addition, crawling these services gives an opportunity to perform a case study of existing tools and platforms.

Finally, the method presented here could be used in other contexts: microtext collections, user lists, and relations could prove useful for microtext corpus building, network visualization, or social network sampling purposes (Gjoka, Kurant, Butts, & Markopoulou, 2011).

Data Sources FriendFeed[13], identi.ca[14] and Reddit[15] were taken into consideration for this study. These services provide a good overview of the peculiarities of social networks. A crawl appears to be manageable by at least the last two of them, in terms of both API accessibility and corpus size, which is not the case for Twitter, for example.

identi.ca identi.ca is a social microblogging service built on open source tools and open standards, which is the reason why I have chosen to crawl it at first. However, access to content was restricted a few months later, so that the methodology described here is not applicable anymore to the website without a user account. This sort of downside concerning web corpora is mentioned by Tanguy (2013).

The advantages compared to Twitter include the Creative Commons license of the content, the absence of limitations on the total number of pages seen (to my knowledge), and the relatively small amount of messages, which can also be a problem. A full coverage of the network where all the information may be publicly available is theoretically possible. Thus, all interesting information is collected and no language filtering is used for this website.

FriendFeed To my knowledge, FriendFeed is the most active of the three microblogging services considered here. It is also the one which seems to have been studied the most by the research community. The service works as an aggregator (Gupta, Garg, Carlsson, Mahanti, & Arlitt, 2009) that offers a broader spectrum of retrieved information. Technically, FriendFeed and identi.ca can overlap, as the latter is integrated in the former. However, the size difference between the two platforms makes this hypothesis unlikely.

[12] http://www.comscore.com/Press_Events/Press_Releases/2011/12/Social_Networking_Leads_as_Top_Online_Activity_Globally

[13] `http://www.friendfeed.com`

[14] `http://www.identi.ca`

[15] `http://www.reddit.com`

The API of FriendFeed is somewhat liberal, as no explicit limits are enforced. Nonetheless, my tests showed that after a certain number of successful requests with little or no sleep, the servers start dropping most of the inbound connections. All in all, the relative tolerance of this website makes it a good candidate to gather a lot of text in a short period of time, even if the time between two requests can vary, as well as the total number of requests per unit of time.

Reddit Reddit is a social bookmarking and microblogging platform, which ranked at 7th place worldwide in the news category according to the site metrics aggregator Alexa at the time this study was conducted. Reddit is now ranked at first place worldwide [16], which makes it a typical Internet phenomenon. The short description of the website according to Alexa is as follows: “User-generated news links. Votes promote stories to the front page." Indeed, the entries are organized into areas of interest called “reddits" or “subreddits". The users account for the linguistic relevance of their channel, the moderation processes are mature, and since the channels (or subreddits) have to be hand-picked, they ensure a certain stability.

Material was gathered for a total of 16 target languages, which can be accessed via so-called “multi-reddit expressions"[17], i.e. compilations of subreddits: Croatian, Czech, Danish, Finnish, French, German, Hindi, Italian, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, and Turkish.

Sadly, it is not possible to go back in time further than the 500th oldest post due to API limitations, which severely restricts the number of links one may crawl. Downloads on a regular basis are necessary to ensure that all possible links are collected. Moreover, requesting the content of the channels regularly makes it possible to get a significant number of links, all the more since the website’s popularity is increasing.

General methodology The methodology used to analyze the URLs is similar to the sketch described above (see p. 151), as it makes use of of the FLUX toolchain.

The only difference between FriendFeed and Reddit on the one hand and identi.ca on the other hand is the spell check performed on the short messages in order to target the non-English ones. Indeed, all new messages on the latter can be taken into consideration, making a selection unnecessary. This spell checking operation is described more in detail above (see p. 154).

Links pointing to media documents, which represent a high volume of links shared on microblogging services, are excluded from the study, as its final purpose is to be able to build a text corpus. As a page is downloaded, links are filtered on the fly using a series of heuristics described below, and finally the rest of the links are stored.

To sum up, the main difference concerning the collection and analysis of URLs published on social networks resides in the collection part, more precisely the fine-tuning of network traversal, link extraction, and link collection processes.

TRUC: an algorithm for TRaversal and User-based Crawls Starting from a publicly available homepage, the crawl engine selects users according to their linguistic relevance based on a language filter, and then retrieves their messages, eventually discovering friends of friends

[16] `http://www.alexa.com/topsites/category/Top/News`
Data from mid 2014.

[17] As an example of multi-reddit expression, here is a possible expression to target Norwegian users: http://www.reddit.com/r/norge+oslo+norskenyheter

and expanding its scope and the size of the network it traverses (see algorithm 2). As this is a breadth-first approach, its applicability depends greatly on the size of the network.

Algorithm 2 Link discovery on the public timeline

```
while there are links do
    if it is not spam or an inappropriate link then
        if the English spell check fails then
            store the link and the user id
        end if
    end if
end while
```

In this study, the goal is to concentrate on non-English speaking messages in the hope of finding non-English links. The main “timeline" fosters a users discovery approach, which then becomes user-centered as the spider focuses on a list of users who are expected to not post messages in English and/or spam. The messages are filtered at each step to ensure relevant URLs are collected. This implies that a lot of subtrees are pruned, so that the chances of completing the traversal increase. In fact, experience shows that a relatively small fraction of users and URLs is selected.

This approach is “static", as it does not rely on any long poll requests (which are, for instance, used to capture a fraction of Twitter’s messages as they are made public); it actively fetches the required pages.

The URLs are discovered using a surface crawl performed on a regular basis that focuses first on the public timeline and then on selected user timelines. Then, a deep miner explores the user network starting from a users list (see algorithm 3). It retrieves all the short messages of a given user. It also performs the same operation with the following users and the followers, but due to the exponential number of requests, not all the messages and users can be retrieved.

The “smart deep crawl" described in the second algorithm and featuring a hostname ratio filter has been implemented, but not thoroughly tested. It is not included in the results section below. It consists in favoring users who mention a variety of hostnames, e.g. whose hostnames-to-total-links ratio on a given page is higher than 10, meaning that if 30 links were retrieved, there have to be more than 3 different hostnames to go deeper into the user’s history.

Check for redirection and sampling Further work on the URL queue before the language identification task ensures an even smaller fraction of URLs really go through the resource-expensive process of fetching and analyzing web documents. The process is as described in section 4.2.1.1: the first step of preprocessing consists of finding those URLs that lead to a redirect, using heuristics or tests. The second step is sampling that reduces both the size of the list and the probable impact of an overrepresented domain name in the result set. If several URLs contain the same domain name, the group is reduced to a randomly chosen URL.

Language identification A similar work on language identification and FriendFeed is described by Celli (2009), who uses a dictionary-based approach: the software tries to guess the language of microtext by identifying very frequent words. However, the fast-paced evolution of the vocabulary used on social networks makes it hard to rely only on lists of frequent terms, therefore my approach seems more complete.

Algorithm 3 Link discovery on user pages

```
function Get the urls
    while there are links do
        if the English spell check fails then
            store the link
            if the user id is new then
                store the user id
            end if
        end if
    end while
    return list of links
end function

while there are user ids do
    fetch the page of a given user
    Get the urls
    if ( different hostnames / total links ) > threshold then
        fetch the next page of a user's timeline
        or fetch a friend's page
        Get the urls
    end if
end while
```

I use both a dictionary-based filter and `langid.py` (see p. 154). The first filter uses a threshold to discriminate between short messages, expressed as a percentage of tokens which do not pass the spell check. The software discriminates between links with mostly English titles and others which are probably in the target language. This option can be deactivated. Tests showed that the probability if finding URLs leading to English text is indeed much higher concerning the "suspicious" list.

Results The surface crawl dealing with the main timeline and one level of depth was performed on all three platforms.[18] In the case of identi.ca, a deep miner was launched to explore the network. FriendFeed proved too large to start such a breadth-first crawler so that other strategies need to be used (Gjoka et al., 2011), whereas the multi-reddit expressions used did not yield enough users.

FriendFeed is the biggest link provider on a regular basis (about 10,000 or 15,000 messages per hour can easily be collected), whereas Reddit is the weakest, as the total figures show.

The total number of English websites may be a relevant indication when it comes to establishing a baseline for finding possibly non-English documents. Accordingly, English accounts for about 55% of the websites, with the second most-used content-language, German, only representing about 6% of the web pages.[19] Consequently, there is a gap between English and the

[18] Several techniques are used to keep the number of requests as low as possible, most notably user profiling according to tweeting frequency. In the case of identi.ca this results in approximately 300 page views every hour.

[19] http://w3techs.com/technologies/overview/content_language/all

other languages, and there is also a discrepancy between the number of Internet users and the content languages.

FriendFeed To test whether the first language filter was efficient, a testing sample of URLs and users was collected randomly. In order to have a reference to compare it to, the first filter was emulated by selecting about 8% of messages (based on a random function) in the spam and media-filtered posts of the public timeline. Indeed, the messages selected by the algorithm approximately amounted to this fraction of the total. At the same time, the corresponding users were retrieved, exactly as described above, and then the user-based step was run. Half of the users' messages was kept, which, according to the real-world data, is realistic.

The datasets compared here were both on the order of magnitude of at least 10^5 unique URLs before redirection checks. At the end of the toolchain, the randomly selected benchmark set comprised 7,047 URLs and the regular set 19,573 URLs.[20] The first was collected in about 30 hours and the second one over several weeks. According to the methodology used, this phenomenon could be explained by the fact that the domain names in the URLs tend to be mentioned repeatedly.

Language	URLs	%
English	4,978	70.6
German	491	7.0
Japanese	297	4.2
Spanish	258	3.7
French	247	3.5

Table 4.4: 5 most frequent languages of URLs taken at random on FriendFeed

According to the language identification system (`langid.py`), the first language filter beats the random function by nearly 30 points (see tables 4.4 and 4.5). The other top languages are accordingly better represented. Other noteworthy languages are to be found in the top 20, e.g. Indonesian and Persian (Farsi).

identi.ca The results of the two strategies followed on identi.ca led to a total of 1,113,783 URLs checked for redirection, which were collected in about a week (the deep crawler reached 37,485 user IDs). A large majority of the 192,327 total URLs apparently led to English texts (64.9%), since only a spam filter was used.

Reddit The figures presented here (see table 4.6) are the results of a single crawl of all available languages together, but regular crawls are needed to compensate for the 500 posts limit. English accounted for 18.1% of the links found on channel pages (for a total of 4,769 URLs) and 55.9% of the sum of the links found on channel and on user pages (for a total of 20,173 URLs).

[20] The figures given describe the situation at the end, after the sampling by domain name and after the selection of documents based on a minimum length. The word URL is used as a shortcut for the web documents they are linked to.

Language	URLs	%
English	8,031	41.0
Russian	2,475	12.6
Japanese	1,757	9.0
Turkish	1,415	7.2
German	1,289	6.6
Spanish	954	4.9
French	703	3.6
Italian	658	3.4
Portuguese	357	1.8
Arabic	263	1.3

Table 4.5: 10 most frequent languages of spell-check-filtered URLs gathered on FriendFeed

Language	URLs	%
English	124,740	64.9
German	15,484	8.1
Spanish	15,295	8.0
French	12,550	6.5
Portuguese	5,485	2.9
Italian	3,384	1.8
Japanese	1,758	0.9
Dutch	1,610	0.8
Indonesian	1,229	0.6
Polish	1,151	0.6

Table 4.6: 10 most frequent languages of URLs gathered on identi.ca

The results in table 4.8 show that the first filter was nearly sufficient for discriminating between the links. Indeed, the microtexts that were under the threshold led to a total of 204,170 URLs. 28,605 URLs remained at the end of the toolchain and English accounted for 76.7% of the documents they were linked to.

The threshold was set at 90% of the words for FriendFeed and 33% for Reddit, each time after a special punctuation strip to avoid the influence of special uses of punctuation on social networks. Yet, the lower filter achieved better results, which could be explained by the moderation system of the subreddits as well as by the greater regularity in the posts of this platform.

Discussion Three main technical challenges had to be addressed, resulting in a separate workflow: the shortened URLs are numerous, yet they ought to be resolved in order to enable the use of heuristics based on the nature of the URLs or a proper sampling of the URLs themselves. The confrontation with the constantly increasing number of URLs to analyze and the necessarily limited resources make website sampling by domain name useful. Finally, the diversity of the web documents rather taxed the language recognition tools, so that a few

Language	URLs	%	Comb. %
English	863	18.1	55.9
Spanish	798	16.7	9.7
German	519	10.9	6.3
French	512	10.7	7.2
Swedish	306	6.4	2.9
Romanian	265	5.6	2.5
Portuguese	225	4.7	2.1
Finnish	213	4.5	1.6
Czech	199	4.2	1.4
Norwegian	194	4.1	2.1

Table 4.7: 10 most frequent languages of filtered URLs gathered on Reddit channels and on a combination of channels and user pages

Language	URLs	% of total
English	21,926	76.7
Spanish	1,402	4.9
French	1,141	4.0
German	997	3.5
Swedish	445	1.6

Table 4.8: 5 most frequent languages of links seen on Reddit and rejected by the primary language filter

tweaks were necessary to correct the results.

The relatively low number of results for Russian can be explained by weaknesses of `langid.py` with deviations of encoding standards. Indeed, a few tweaks were necessary to correct the biases of the software in its pre-trained version, in particular regarding texts falsely considered as being written in Chinese, although URL-based heuristics indicate that the website is most probably hosted in Russia or Japan. A few charset encodings found in Asian countries were also a source of classification problems. The low-confidence responses as well as a few well-delimited cases were discarded in this study, as they account for no more than 2% of the results. Ideally, a full-fledged comparison with other language identification software would be necessary to identify its areas of expertise.

Similarly to the study on low-resourced languages,cloaking has not been addressed so far.

Regarding topics, a major user bias was not addressed either. Among the most frequently shared links on identi.ca, for example, are those related to technology, IT, or software, and mostly written in English. The social media analyzed here tend to be dominated by English-speaking users, either native speakers or second-language learners.

In general, there is room for improvement concerning the first filter. The threshold could be tested and adapted to several scenarios. This might involve larger datasets for testing purposes and machine learning techniques relying on feature extraction.

The contrasted results on Reddit shed a different light on the exploration of user pages: in all likelihood, users mainly shared links in English if they are not posting them on a language-relevant channel. The results on FriendFeed were better from this point of view, which might

suggest that English is not used equally on all platforms by users who speak languages other than English. Nonetheless, there seemed to be a strong tendency for the microblogging services discussed here to be mainly English-speaking.

Last but not least, the adequacy of the web documents shared on social networks has yet to be thoroughly assessed. Other fine-grained instruments (Schäfer & Bildhauer, 2012) as well as further decisions (Schäfer et al., 2013) are needed along the way from the output of this toolchain to a full-fledged web corpus.

Conclusion I have presented a methodology for gathering multilingual URLs on three microblogging platforms. In order to do so, I performed traversals of the platforms and used already available tools to filter the URLs accordingly and to identify their language.

I have provided open source software to access the APIs (FriendFeed and Reddit) and HTML version of identi.ca, as authentication is mandatory for the API. The TRUC algorithm is fully implemented. All the operations described in this paper can be reproduced using the same tools, which are part of repositories currently hosted on GitHub.[21]

The main goal was achieved, as hundreds, if not thousands, of URLs for lesser-known languages such as Romanian or Indonesian could be gathered on social networks and microblogging services. When it comes to filtering out English posts, a first step using an English spell checker gave better results than the baseline established using microtexts selected at random. However, the discrepancy was remarkable between the languages one would expect to find based on demographic indicators on the one hand, and the results of the study on the other hand. English websites stayed numerous even when one tried to filter them out.

This proof of concept is usable, but a better filtering process and longer crawls may be necessary to unlock the full potential of this approach. Lastly, a random-walk crawl using these seeds and a state of the art text categorization may provide more information on what is really shared on microblogging platforms.

Future work perspectives could include dealing with live tweets (as Twitter and FriendFeed can be queried continuously), exploring the depths of identi.ca and FriendFeed, and making the directory of language-classified URLs collected during this study publicly available.

4.2.1.4 Comparison of available sources

Looking for alternatives, what issues do we face? Search engines actually yield good results in terms of linguistic quality. Besides, it is not possible to start a web crawl from scratch, so the main issue to be tackled can be put this way: where may we find web pages bound to be interesting for corpus linguists, and which in turn contain many links to other interesting web pages?

Researchers in the machine translation field have started another attempt to outsource competence and computing power, making use of data gathered by the CommonCrawl project[22] to find parallel corpora (Smith et al., 2013). Nonetheless, the quality of the links may not live up to their expectations. A series of issues detailed above (see above) hinders the collection of high-quality language samples. The most important factors can be listed as follows.

First, purely URL-based approaches are a trade-off in favor of speed while sacrificing precision, and language identification tasks are a good example of this phenomenon (Baykan et

[21] https://github.com/adbar/microblog-explorer
[22] http://commoncrawl.org/

al., 2008). Second, machine-translated content is a major issue, as is text quality in general, especially when it comes to web texts (Arase & Zhou, 2013). Third, mixed-language documents slow down text gathering processes (King & Abney, 2013). Fourth, link diversity is a also problem, which in my opinion has not got the attention it deserves. Last, the resource is constantly moving. There are not only fast URL changes and ubiquitous redirections. Following the "web 2.0" paradigm, much web content is being injected from other sources, so that many web pages are now expected to change at any time.[23] Regular exploration and re-analysis could be the way to go to ensure the durability of the resource.

In this subsection, I have introduced a scouting approach which considered the first issue, touched on the second one, provided tools and metrics to address the third and fourth, and adapted to the last. In the following section I will describe my methodology, then I will show in detail which metrics I decided to use, and last I will discuss the results.

Languages studied I have chosen four different languages in order to see if my approach generalizes well: Dutch, French, Indonesian and Swedish. This enables me to compare several language-dependent web spaces which ought to have different if not incompatible characteristics. In fact, the "speaker to website quantity" ratio is probably extremely different when it comes to Swedish and Indonesian. I showed in a previous study that this greatly affects link discovery and corpus construction processes (Barbaresi, 2013a).

French is spoken on several continents and Dutch is spoken in several countries (Afrikaans was not part of this study). Indonesian offers an interesting point of comparison, as the chance to find web pages in this language during a crawl at random is scarce. For this very reason, I explicitly chose not to study English or Chinese because they are clearly the most prominently represented languages on the web.

Data sources I used two reference points, the first one being the existing method depending on search engine queries, upon which I hoped to cast a new light with this study. The comparison is based on URLs retrieved using the BootCaT seed method on the meta-engine E-Tools[24] at the end of 2012.[25] The second reference point consisted of social networks, to whose linguistic structure I already dedicated a study (Barbaresi, 2013b) where the method used to find the URLs is described in detail. I chose to adopt a different perspective, to re-examine the URLs I gathered and to add relevant metadata in order to see how they compared to the other sources studied here.

I chose to focus on the three different networks which I analyzed in detail in the study above: FriendFeed, an aggregator that offers a broader spectrum of retrieved information; identi.ca, a microblogging service similar to Twitter; and Reddit, a social bookmarking and microblogging platform. Perhaps not surprisingly, these data sources display the issues linked to API instability mentioned above. The example of identi.ca mentioned in the study above is telling: until March 2013, when the API was closed after the company was bought up, it was a social microblogging service built on open source tools and open standards. The advantages compared to Twitter included the Creative Commons license of the content, and the absence of limitations on the total number of pages seen.

[23] This is the reason why Marco Baroni states in the talk mentioned above that his "love affair with the web" is over.

[24] http://www.etools.ch/

[25] Thanks to Roland Schäfer for letting me use the URLs extracted from E-Tools and DMOZ.

Another data source was the Open Directory Project (DMOZ[26]), where a selection of links is curated according to their language and/or topic. The language classification was expected to be adequate, but the amount of viable links was an open question, as was the content.

Last, the quality of links on the free encyclopedia Wikipedia was expected to be high. It is well established that this encyclopedia is a useful resource in a given language edition. The open question resided in the links pointing to the outside world, as it is hard to get an idea of their characteristics due to the large number of articles, which is rapidly increasing even for an under-resourced language such as Indonesian.

Processing pipeline and metadata The processing pipeline was the FLUX toolchain, described above (see above). It consists of visits to a series of websites in order to compute useful metadata which are then made available together with a selection of URLs.

The metadata described in this subsection can be used in classificatory or graph-based approaches. I have used some of them in the results below but did not exhaust all the possible combinations in this study. There are nine of them in total, which can be divided into three categories: corpus size metrics, which are related to word count measures, web science metrics, which ought to be given a higher importance in web corpus building, and finally language identification, performed using an external tool.

Corpus size metrics Web page length (in characters) was used as a discriminating factor. Web pages which are too short (less than 1,000 characters long after HTML stripping) were discarded in order to avoid documents containing just multimedia (pictures and/or videos) or microtext collections for example, as the purpose was to simulate the creation of a general-purpose text corpus.

The page length in characters after stripping was recorded, as well as the number of tokens, so that the total number of tokens of a web corpus built on this URL basis can be estimated. The skewed distribution of the length of web pages was also a hint at duplicate content and consequently a sinking quality of the document collection (see above).

Web science metrics Similarly to the two studies above, host sampling is a very important step because the number of web pages is drastically reduced, which makes the whole process more feasible and more well-balanced, i.e. less prone to host biases. The present study gives evidence in the form of IP-based statistics which corroborate this hypothesis, as shown below.

The deduplication operation is elementary, it takes place at document level, using a hash function. The IP diversity is partly a relevant indicator, as it can be used to prove that not all domain names lead to the same server. Nonetheless, it cannot detect the duplication of the same document across many different servers with different IPs, which in turn the elementary deduplication is able to reveal.

Links leading to pages within the same domain name and links leading to other domains were extracted from the HTML markup. The first figure can be used to find possible spam or irrelevant links, with the notable exception of websites like Amazon or Wikipedia, which are quite easy to list. The latter may be used to assess the richness (or the suspiciousness) of a website by the company it keeps. While this indicator is not perfect, it enables users to draw conclusions without fetching all the downstream URLs.

[26] http://www.dmoz.org/

Language identification Using a language identification system has a few benefits: it enables finding “regular" texts in terms of statistical properties as well as excluding certain types of irregularities such as encoding problems. Web text collections are smoothed out in relation to the statistical model applied for each language target, which is a partly destructive but interesting feature.

There are cases where the confidence interval of the language identifier is highly relevant, for instance if the page is multi-lingual. There are two main effects in that case: on the one hand the confidence indicator gets a lower value, so that it is possible to isolate pages likely to be in the target language only. On the other hand, the language guessed at is the one with the largest number of identifiable words: if a given web page contains 70% Danish and 30% English, then it will be classified as being written in Danish, with a low confidence interval: this information is part of the metadata I associated with each web page. Since nothing particular stood out in this respect I will not mention it further.

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
Dutch	12,839	1,577	84.6	27,153	3,600	5,325,275	73.1
French	16,763	4,215	70.2	47,634	8,518	19,865,833	50.5
Indonesian	110,333	11,386	66.9	49,731	8,634	50,339,311	18.6
Swedish	179,658	24,456	88.9	24,221	9,994	75,328,265	20.0

Table 4.9: URLs extracted from search engines queries

Results: characteristics of the BootCaT approach First of all, I let my toolchain run on URLs obtained using the BootCaT approach, in order to get a glimpse of its characteristics. I let the URL extractor run for several weeks on Indonesian and Swedish and only a few days for Dutch and French, since I was limited by the constraints of this approach, which becomes exponentially slower as one adds target languages.[27] The results discussed below are displayed in Table 4.9.

The domain name reduction had a substantial impact on the set of URLs, as about a quarter of the URLs at best (for French) had different domain names. This was a first hint at the lack of diversity of the URLs found using the BootCaT technique.

Unsurprisingly, the majority of links appeared to be in the target language, although the language filters did not seem to perform very well. As the adequate matching of documents to the user’s language is paramount for search engines, it is probably a bias of the querying methodology and its random tuples of tokens. In fact, it is not rare to find unexpected and undesirable documents such as word lists or search engine optimization traps.

The length of web documents was remarkable, it indicated that they were likely to contain long texts. Moreover, the median length seemed to be quite constant across the three languages at about 8,000 tokens, whereas it was less than half that (3,600) for Dutch. All in all, it appeared to be an advantage which clearly explained why this method is considered to be successful. The

[27]The slow URL collection is explained by the cautious handling of this free and reliable source, implying a query rate limiting on my side. The scouting approach by itself is a matter of hours.

potential corpus sizes are noteworthy, especially when enough URLs are gathered in the first place, which was already too impracticable in my case to be considered a sustainable option.

The number of different IPs, i.e. the diversity in terms of hosts, seemed to gradually lower as the URL list became larger. The fact that the same phenomenon occurred for Indonesian and Swedish, with one host out of five being “new", indicates a strong tendency.

	% in target	URLs retained	Length		Tokens (total)	Different IPs (%)
			mean	median		
Dutch	0.6	465	7,560	4,162	470,841	68.8
French	5.9	4,320	11,170	5,126	7,512,962	49.7
Indonesian	0.5	336	6,682	4,818	292,967	50.9
Swedish	1.1	817	13,807	7,059	1,881,970	58.5

Table 4.10: URLs extracted from a blend of social networks crawls (FriendFeed, identi.ca, and Reddit) with no language target. 738,476 URLs analyzed, 73,271 URLs retained in the global process.

Results: Social networks Due to the mixed nature of the experimental setting, no conclusions can be drawn concerning the single components. The more than 700,000 URLs that were analyzed give an insight regarding the usefulness of these sources. About a tenth of them remained as responding websites with different domain names, which is the lowest ratio within this study. This could be explained by the fast-paced evolution of microblogs and also by the potential impurity of the source compared to the user-reviewed directories whose results I will describe next.

As I did not target the studied languages during the URL collection process, there were merely a few hundred different domain names to be found, with the exception of French, which was a lot more prominent.

Table 4.10 provides an overview of the results. The mean and median lengths are clearly lower than in the search engine experiment. In the case of French, with a comparable number of remaining URLs, the corpus size estimate is about 2.5 times smaller. The host diversity is comparable, and does not seem to be an issue at this point.

All in all, social networks are probably a good candidate for web corpora, but they require a focused approach to microtext in order to target a particular community of speakers.

Results: DMOZ As expected, the number of different domain names on the Open Directory project was high, giving the best ratio in this study between unfiltered and remaining URLs. The lack of web pages written in Indonesian was a problem for this source, whereas the other languages seemed to be far better covered. The adequacy of the web pages with respect to their language was excellent, as shown in Table 4.11. These results underline the quality of the resource.

On the other hand, document length is the biggest issue here. The mean and median values indicate that this characteristic is quite homogeneous throughout the document collection. This may easily be explained by the fact that the URLs which are listed on DMOZ mostly lead to

corporate homepages for example, which are clear and concise, the eventual “real" text content being somewhere else. What’s more, the websites in question are not text reservoirs by nature. Nonetheless, the sheer quantity of listed URLs compensates for this fact. The corpus sizes for Dutch and French are quite reasonable if one bears in mind that the URLs were sampled.

The relative diversity of IPs compared to the number of domain names visited was another indicator that the Open Directory leads to a wide range of websites. The directory performed well compared to the sources mentioned above, it was also much easier to crawl. It did not cost us more than a few lines of code followed by a few minutes of runtime to gather the URLs.

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
DMOZ							
Dutch	86,333	39,627	94.0	2,845	1,846	13,895,320	43.2
French	225,569	80,150	90.7	3,635	1,915	35,243,024	33.4
Indonesian	2,336	1,088	71.0	5,573	3,922	540,371	81.5
Swedish	27,293	11,316	91.1	3,008	1,838	3,877,588	44.8
Wikipedia							
Dutch	489,506	91,007	31.3	4,055	2,305	15,398,721	43.1
French	1,472,202	201,471	39.4	5,939	2,710	64,329,516	29.5
Indonesian	204,784	45,934	9.5	6,055	4,070	3,335,740	46.3
Swedish	320,887	62,773	29.7	4,058	2,257	8,388,239	32.7

Table 4.11: URLs extracted from DMOZ and Wikipedia

Results: Wikipedia The characteristics of Wikipedia are quite similar, since the free encyclopedia also makes dumps available, which are easily combed through in order to gather start URLs. Wikipedia also compares favorably to search engines or social networks when it comes to the sampling operation and page availability. It is a major source of URLs, with numbers of gathered URLs in the millions for languages like French. As Wikipedia is not a URL directory by nature, it is interesting to see what the characteristics of the pages it links to are. The results are shown in Table 4.11.

First, the pages referenced in a particular language edition of Wikipedia often pointed to web pages written in a different language. According to my figures, this was a clear case, all the more since web pages in Indonesian are rare. Still, with a total of more than 4,000 retained web texts, it fared a lot better than DMOZ or social networks.

The web pages were longer than the ones from DMOZ, but shorter than the rest. This may also be related to the large number of concise homepages in total. Nonetheless, the impressive number of URLs in the target language is decisive for corpus building purposes, with the second-biggest corpus size estimate obtained for French.

The IP-related indicator yielded good results with respect to the number of URLs retrieved. Because of the high number of analyzed URLs the figures between 30 and 46% give an insight into the concentration of web hosting providers on the market.

Language	Min.	Max.	Average
Search Engines			
fr	0	25,028	33.80
id	0	22,912	20.28
nl	0	10,332	29.26
sv	0	50,004	88.44
DMOZ			
fr	0	5,989	8.47
id	0	497	15.80
nl	0	1,907	6.38
sv	0	361	6.11
Wikipedia			
fr	0	4,563	12.82
id	0	5,831	15.17
nl	0	2,399	11.13
sv	0	5,059	11.69
Social networks			
all	0	10,908	18.85

Table 4.12: Comparison of the number of different domains in the page's outlinks for four different sources

Distribution of outlinks I also analyzed the results regarding the number of links that led out of the page's domain name. The distribution of links pointing to websites located out of the original domain is summarized in Table 4.12. Out-of-domain links (sometimes also called outlinks in the literature) can be used as an indicator estimating the potential richness of a crawl. In fact, when crawling on a large-scale basis, one usually expects to gather a wide array of different websites, because this potentially leads to a more diverse content in terms of genres or language uses, as well as to a broader spectrum of users.

That is why the number of outlinks can be useful in order to select web pages based on their potential. However, the estimation is not always reliable, due to numerous crawler traps and deception techniques related to search-engine-optimization (see above). It relies on the extraction and comparison of the domain names in the final URL (after possible redirections) and in each of the web page's links. Only different domain names count.

As the maximum values in Table 4.12 show, the number of outlinks can be absurdly high. From a technical point of view, a web page pointing to 20,000 different domain names is almost certainly a scam. Thus, the values contained in the metadata should be taken with a pinch of salt, and filtered according to the research goals.

The average values are no surprise, since the average number of different domains linked in a single web page is noticeably higher concerning search engine results. As a matter of fact, this

number is a discriminating tool used by search engines to rank pages. From that standpoint, the numbers mentioned here seem to validate the precision of the chosen methodology.

The results were rather consistent across the languages studied here, the main differences existing between the different sources. In fact, there seemed to be a tendency towards a hierarchy in which the search engines are on top, followed by social networks, Wikipedia and DMOZ. This is one more hint at the heterogeneous nature of the data sources I examined with respect to the criteria I chose.

Discussion The hierarchy in terms of the outlinks is one more reason why search engines queries are believed to be fast and reliable in terms of quantity. This method was fast, as the web pages are long and full of links, which enables rapid harvesting of a large number of web pages without having to worry about going round in circles. The researchers using the BootCaT method probably took advantage of the undocumented but efficient filtering operations which search engines perform in order to find reliable documents. Since this process takes place in a competitive sector where this kind of information can be sold, it may explain why the companies now try to avoid giving it away for free.

In the long run, several questions regarding URL quality remain open. As I showed using a high-credibility source such as Wikipedia, the search engines results are probably closer to the maximum amount of text that is to be found on a given website than the other sources, all the more when the sampling procedure chooses a page at random without analyzing the rest of a website and thus without maximizing its potential in terms of tokens. Nonetheless, confrontation with the constantly increasing number of URLs to analyze and necessarily limited resources make a website sampling by domain name useful.

This was part of my cost-efficient approach, where the relatively low performance of Wikipedia and DMOZ is compensated by the ease of URL extraction. Besides, the size of the potential corpora mentioned here could increase dramatically if one was to remove the domain name sampling process and if one was to select the web pages with the most out-domain links for the crawl.

What's more, DMOZ and Wikipedia are likely to improve over time concerning the number of URLs they reference. As diversity and costs (temporal or financial) are real issues, a combined approach could take the best of all worlds and provide a web crawler with distinct and distant starting points, between the terse web pages referenced in DMOZ and the expected "freshness" of social networks. This could be a track to consider, as they could provide a not inconsiderable amount of promising URLs.

Finally, other fine-grained instruments as well as further decisions processes (Schäfer et al., 2013) will be needed on the way from the output of the toolchain to a full-fledged web corpus. The fact that web documents coming from several sources already differ by our criteria does not exclude further differences regarding text content. By way of consequence, future work could include a few more linguistically relevant text quality indicators in order to go further in bridging the gap between web data, NLP and corpus linguistics.

Conclusion I have evaluated several strategies for finding texts on the web. The results established no clear winner, so complementary approaches are called for. In light of these results, it seems possible to replace or at least to complement the existing BootCaT approach. It is understandable why search engine queries have been considered a useful data source. However, I

have shown that they lack diversity at some point, which apart from their impracticality may provide sufficient impetus to look for alternatives.

4.2.1.5 General conclusion on the linguistic seed directory

Approach I have discussed how I address several issues in order to design robust processing tools which (combined with the diversity of sources and usable metadata) enable researchers to get a better glimpse of the course a crawl may take. The problem of link diversity has not been well-studied in the context of corpus linguistics; I have presented metrics to help quantify it and showed a possible route for gathering a corpus using several sources leading to a satisfying proportion of different domain names and hosts.

The metadata collected in the studies can be used to draw a partial view of the observed network. Consequently, web page metrics regarding the importance, weight, or quality of content can be estimated, for instance using the PageRank algorithm.[28]

Moreover, even if I did not take advantage of this information for this study, the fetcher also records all the links it "sees" (as an origin-destination pair), which enables graph-based approaches such as visualization of the gathered network or the assessment of the "weight" of a website in the URL directory. Also, this metadata may very well be useful for finding promising start URLs.

Overall statistics The database consisting of the output of the processing chain (see p. 152) is named LSD, which stands for Language-classified Seed Directory.

It is meant to be a large database of pre-classified URLs seeds which can be used as start URLs for crawlers, as a complement or a replacement for the BootCaT method.

As of as of March 2013, the following information were stored in the directory:

- hash value of the URL
- language of the document and confidence indicator
- length before / after HTML stripping
- number of words
- number of in- and outlinks
- IP address(es) of the host
- content of the HTTP-Last-Modified field (if any)

The overall database statistics as of March 2013 were the following:

- Total URLs seen: 20,032,862
- Total URLs FLUXed: 2,484,095
- Number of different IPs: 683,900
- Mean / Median number of characters: 2198 / 4949

[28] `https://en.wikipedia.org/wiki/PageRank`

- Mean / Median number of words: 282 / 652
- Mean in- / outlinks: 9.8 / 10.1
- Estimated median freshness of documents: Sept 18, 2012

The relatively high number of different IPs gathered in a short period, since FLUX has been mainly used for comparison of sources and rarely as a crawler, shows the potential of a combination of, on the one hand, a selection of high-quality sources, and on the other hand a light scout approach.

The rest of the statistical information highlighted that the pages are relatively short, which may be explained by the influence on the total of the studies performed on DMOZ, but also relatively fresh, i.e. about six months old in average, which was encouraging and demanding at the same time, because it indicated that it may be necessary to examine the web pages on a regular basis in order to maintain current results.

4.2.2 Impact of prequalification on (focused) web crawling and web corpus sampling

4.2.2.1 Introduction and definitions

So-called focused crawlers (in a broad sense) are designed to maximize the weighted coverage (Olston & Najork, 2010) with respect to some specific definition of document weight, for example when documents with a high search-engine relevance (measured as its Page-Rank or a similar score), documents about specific subjects, or documents in a specific language are desired.

Concerning web corpus crawling, a document with a high weight can simply be defined as one which is not removed from the corpus by the post-processing tools due to low linguistic quality and/or a document which contributes a high amount of text to the corpus.

More precisely, in the case of linguistic focusing, the ideal would be to find good corpus documents early and sustain a high rate throughout the crawl.[29]

Figure 4.2 exemplifies a phenomenon which could be termed the “exhaustion" of the seeds. Indeed, the quality drops during the crawl.

Recently, an interesting approach to crawl optimization along such lines was suggested which relies on statistics about the corpus yield from known hosts (Suchomel & Pomikálek, 2012). Under this approach, the weight (rather of a whole web host) is taken to be the ratio of good documents from the host remaining in the corpus after a specific post-processing chain has been applied to the documents. Harvested URLs pointing to certain hosts are prioritized accordingly.

Together with Roland Schäfer and Felix Bildhauer, I have conducted work at the FU Berlin regarding optimization of crawling processes (Schäfer, Barbaresi, & Bildhauer, 2014). We follow a similar route as Suchomel and Pomikálek (2012), but look at document-local features instead of host statistics. We choose to refer to weighted coverage as “yield ratio", as they are related notions. We define the yield ratio Yd for a set Dc of crawled unprocessed documents and

[29]The work presented here is the result of joint experiments with Felix Bildhauer and Roland Schäfer at the FU Berlin. Unless marked otherwise, the points I describe in detail are essentially my contributions to the web corpus project.

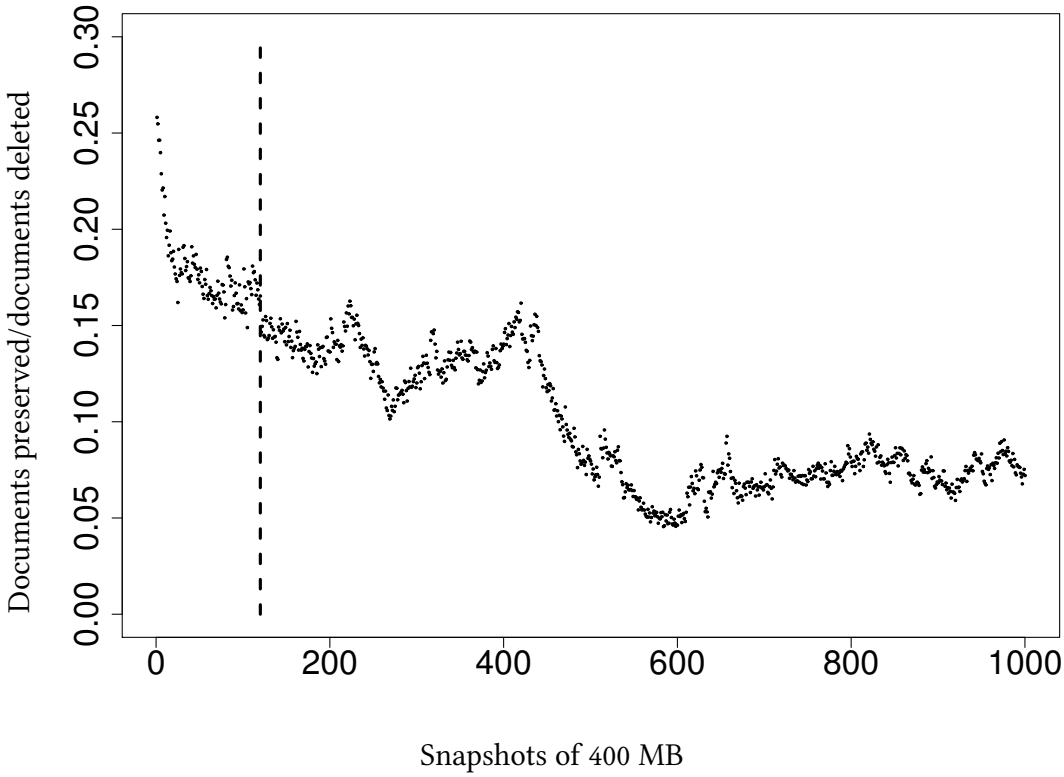


Figure 4.2: Ratio of documents preserved vs. deleted over a breadth-first crawl of 390 GB with many "high-quality" seed URLs are snapshots of 400 MB; the vertical line marks the point of seed URLs exhaustion. (Schäfer & Bildhauer, 2013, p. 31)

a set Dr of retained documents after filtering and processing for inclusion in a corpus, with $Dr \subset Dc$, as:

$$Yd = |Dr|/|Dc|$$

For example, a document yield ratio $Yd = 0.21$ means that 21% of the crawled documents survived the cleaning procedure (i.e. were not classified as duplicates or spam, were long enough, written in the target language, etc.) and ended up in the corpus. In order to maximize Yd, 79% of the documents should not have been downloaded in the first place in this example. A parallel definition is assumed for Yb for the respective amounts of bytes. The document yield ratio is easier to interpret because the byte yield ratio depends on the amount of markup which has to be stripped, and which might vary independently of the quality of the downloaded web pages.

Obviously, the yield ratio – like the weighted coverage – depends highly on the definition of what a good document is, i.e. what the goal of the crawl is. We assume, similarly to the approach of Suchomel and Pomikálek (2012), that our tools reliably filter out documents that are interesting documents for inclusion a corpus, and that calculating a yield ratio based on the output of those tools is therefore reasonable.

4.2.2.2 Experiment: Seed and Crawl Quality

Setting In the resulting experiment, also documented in Schäfer et al. (2014), we examine the correlation between the yield ratio of crawler seed URLs and the yield ratio of short Breadth-First Search (BFS) crawls based on those URLs. We used the Heritrix (version 1.14) web crawler (Mohr, Stack, Rnitovic, Avery, & Kimpton, 2004) and an older version of the texrex web page cleaning toolkit (Schäfer & Bildhauer, 2012). The tools perform, among other things, boilerplate detection and text quality evaluation in the form of the so-called Badness score (Schäfer et al., 2013).

A document receives a low Badness score if the most frequent function words of the target language have a high enough frequency in the document (see next section). This experiment was carried out in the context of an evaluation of sources of different seed URLs for crawls (see above), concerning Dutch, French, Indonesian and Swedish.

We randomly sampled 1,000 seed URLs for each of the 20 permutations of seed sources

and languages/TLDs, downloaded them and used texrex to determine the document yield ratio for the documents behind the 1,000 seeds. The software was configured to perform boilerplate removal, removal of documents based on high Badness scores, perfect duplicate removal, and deletion of documents shorter than 1,000 characters (after boilerplate removal). Then, we crawled the respective TLDs, starting the crawls with the 1,000 seed URLs, respectively. In each crawl, we downloaded 2 GB of raw data, cleaned them, and calculated the document yield ratio using the same configuration of texrex as we used for cleaning the seed documents. Figure 4.3 plots the data and an appropriate linear model.

Results Figure 4.3 exemplifies that there is a strong correlation (adjusted $R2 = 0.7831$) between the yield ratio of the documents behind the seed URLs and the yield ratio of the documents found by using the seeds for BFS crawling. It follows that giving high priority to links from pages which are themselves considered high-quality documents by the post-processing tools will likely lead to more efficient crawling. Since there is no fundamental distinction between initial URL seeds and URLs harvested at a later time during the crawl, this effect is likely to extend to the whole run time of a crawl.

Additionally, figure 4.4 summarizes the results of an experiment on qualification involving a combination of sources (DMOZ, etools (meta search engine), Friendfeed, and identi.ca) and languages (Dutch, French, Indonesian, and Swedish). Short breadth-first crawls, each time a few hours long, were started based on URLs analyzed by FLUX, then the web documents were processed as if a general-purpose corpus with general settings was built. Figure 4.4 depicts the correlation between the source and the quality of all documents collected during the crawl phase.

The overall quality is rather good. The most striking and, for us, counter-intuitive result is the fact that document quality is a function of languages rather than sources. The documents in French for instance are all considered to be of slightly inferior quality by the software. Another conclusion to draw is that document quality can actually improve during the beginning of the crawl, as it seems to be the case for most sources, search engines and identi.ca.

The quality of other sources after the filtering by FLUX seems to outweigh search engines in most cases, concerning these short experiments. In this study, the social networks identi.ca and Friendfeed seem to be noteworthy alternatives to search engine queries.

Conclusion All in all, we have shown in Schäfer et al. (2014) that two standard cleaning algorithms used in web corpus construction, i.e. text quality evaluation based on frequent short

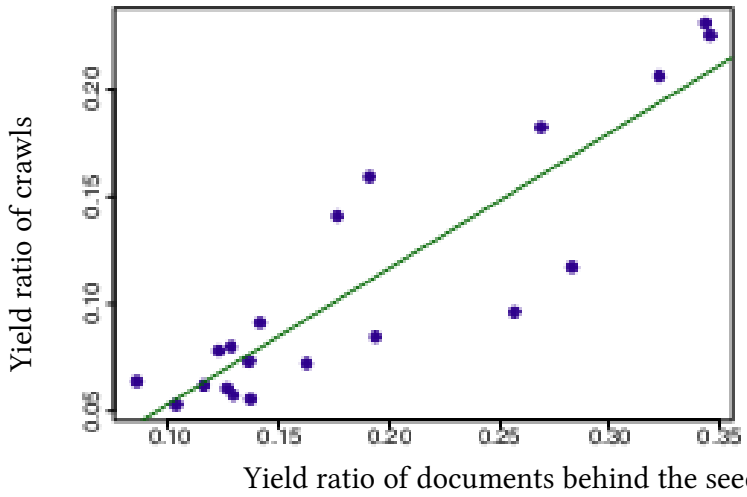


Figure 4.3: Yield ratio Yd of the crawls (y axis) plotted against the yield ratio of the documents behind the crawls' 1,000 seeds (x axis). (Higher Yd is better.) Linear model: $Intercept = -0.0098$, $Coefficient = 0.6332$, $R2 = 0.7831$ (adjusted), $p < 0.001$ (ANOVA). (Schäfer, Barbaresi, & Bildhauer, 2014)

words and boilerplate detection, have a high potential for optimizing web corpus crawling through the prioritization of harvested URLs in a crawler system. In that sense, a light scout collecting links before a massive crawl may not only improve the crawl by letting it take a more fruitful direction, but also gain time, as less unwanted documents are being downloaded and processed.

4.2.3 Restricted retrieval

4.2.3.1 Processing chain

URL-based (restricted) retrieval Although there are distinctions to be made between potential crawling strategies (see above), what I have called restricted retrieval here is roughly comparable to focused crawling.

In fact, it is a hybrid approach which benefits from manual intervention. The crawl does not merely follow the instructions of the robots.txt files (see above), it also uses a blacklist which can be defined manually, in order to avoid particular (unwanted) types of content, such as cooking recipes on a newspaper website or videos.

Example: extracting list of links If the website has an archive, a sitemap or a general list of its contents one can save time by picking the interesting links once and for all. Algorithm 4 describes a basic way to crawl a particular website.

Download en masse (archiving) or on-the-fly extraction (scraping) Once the list of target URLs is set, the download can begin. At that point, one may distinguish two main types of retrieval.

In the first, an en masse download of the target URLs and everything they link to within the same domain name is performed. This approach is very close to the one adopted by web archives, as so-called dumps or mirrors of a whole web site are made.

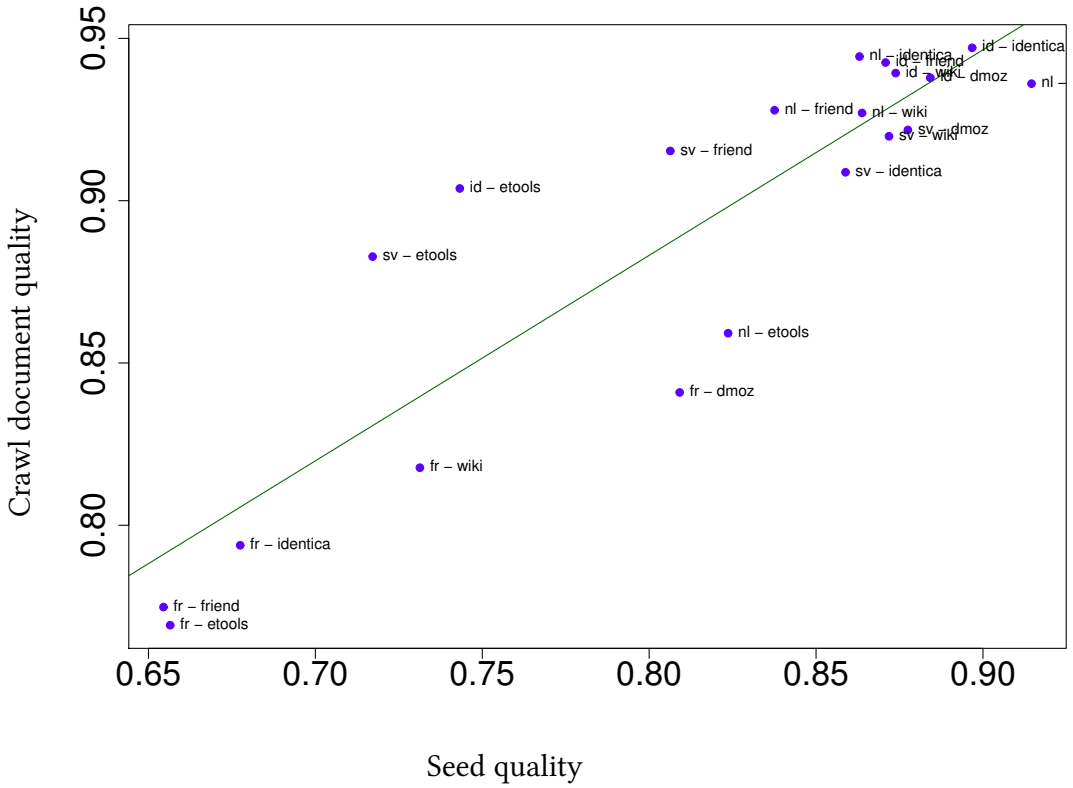


Figure 4.4: LM for crawl document quality modeled by seed quality on short breadth-first crawls ($R^2 = 0.79$). Four different languages: Dutch (nl), French (fr), Indonesian (id), and Swedish (sv). Four different sources: DMOZ, etools (meta search engine), Friendfeed, and identi.ca. Source: joint work with F. Bildhauer and R. Schäfer.

There are several caveats to such an approach. Contrary to the procedure for web archives, not all the content is to be stored, but only relevant text, meaning a fraction. The download operation requires storage space, sometimes quite a lot, even though part of the downloaded material is not strictly necessary. Additionally, as duplicate content is frequent on the Web, it may be that documents have to be filtered in this respect, thus making the use of storage space even more questionable.

However, as the different filtering operations leading to the final text collection may change in the course of time, it could be useful to have a full archive to start from, in order to ensure reproducibility of filtering and make potential comparisons possible.

In the second case, content is extracted on the fly, which is sometimes named web scraping. It is then impossible to come back to a “raw" version of the sourced web pages. That is why the scraping process has to be optimized before being used on a large scale, as results will not be comparable otherwise.

On-the-fly extraction: example of a page explorer The main scraping component is the module that indexes or in this particular case selects the desired content and stores it in a file. The algorithm 5 describes a potential way to write an program to perform scraping.

The scraping approach is also feasible regarding web crawling by itself, meaning that one

Algorithm 4 Example of algorithm for restricted crawling

```
while going through a shortlist of archives do
    fetch page
    extract the links
    for each link do
        if the link matches a given regular expression then
            store in a list
        end if
    end for
    if there are other results or archives pages then
        for all the available items do
            add ?p=... or ?page=... or anything suitable to the last seen page
            fetch page, find the links, process each link
        end for
    end if
end while
remove the duplicate items from the list
write to file (after a last control)
```

Algorithm 5 Example of algorithm for web scraping

```
while going through the list of pages do
    fetch page
    cut the top and the bottom (e.g. HTML HEAD and unwanted information)
    if there is something like a <div class="title"> or <h1> then
        extract it
        clean it
        store it
    end if
    look for the paragraphs of the text, clean and store them
end while
write the text with desired meta-information to file
```

may also scrape the links on the fly. It lacks the systematic approach and the clear overview which is provided by a precomputed URL list, but it is potentially efficient on every website. Moreover, it is interesting to see how far one may go starting from one single page for each given website.

If there are several pages, one can either change the URL before the page is fetched (if there is a “text on one page" option) or one can proceed as described by the links.

Metadata extraction and completion Metadata extraction is part of boilerplate removal, since meta-information contained for instance in the header or in particular HTML tags should be separated from the actual content. At the same time, relevant information can be extracted inside and outside of the document, external sources can help to complete metadata extraction. The URL itself is relevant, as well as download information such as time indications and details

in server communication.

Especially concerning news articles, the contents of a given website are prone to change daily if not hourly. For this reason it is advisable to explore the website bit by bit (or more precisely rubric by rubric) and/or on a regular basis, rather than to perform a full-depth crawl once only.

One might want to write a shell script to fire the two main phases automatically and/or to save the results on a regular basis. That way, if something goes wrong after a reasonable amount of explored pages not all the work is lost.

4.2.3.2 Examples of several approaches to retrieval

Example 1: Sources of political speeches The speeches were crawled from the online archive of the German Presidency (bundespraesident.de) and from the official website of the German Chancellery (bundesregierung.de). Strange as it seems, the resource is not stable, with texts appearing or disappearing for no obvious reason. Additionally, not all the speeches in the corpus can be found on this website anymore due to a change of design. Further details are mentioned in Barbaresi (n.d.).

This ordering was made using regular expressions in both titles and excerpts. It seemed to work properly but does not guarantee a perfect classification.

An automaton stripped off the salutatory addresses of the speeches using regular expressions, with good accuracy, although not perfect due to the extreme variation among speakers.

In the texts from before 2005 the encoding is deficient, mostly affecting the punctuation marks and the spaces. This is a typical issue for corpora from the web, which is still easier to deal with than OCR problems.

Example 2: Crawling of newspaper websites Starting from the front page or from a given list of links, the crawlers retrieve newspaper articles and gather new links to explore them as they go. The HTML code as well as superfluous text are stripped in order to save disk space, the remaining text (which is possibly the exact content of the article) is saved as a raw text file with relevant metadata (such as title, subtitle, excerpt, author, date and URL). Details are described in ? (?).

The crawlers also detect and filter undesirable content based on a URL analysis, as the URLs give precious hints about the article column or about website specifics and internal documents. In fact, a few columns were discarded because most of the text they contained was of a different nature or genre, such as the cooking recipes of *Die Zeit* or the pictures of topless women on the website of *Bild*.

As the crawling process took place at a time when online versions of the newspapers were emerging, it was impacted by the editorial changes related to this evolution. The duplication of the articles (see below) as well as the existence of near duplicate documents are symptomatic for the erratic decisions taken by the editorial staff, leading for example to the parallel existence of online and print versions on the website of *Die Zeit* in 2008.

Example 3: OpenSubtitles The subtitles were retrieved from the OpenSubtitles project[30], a community-based web platform for the distribution of movie and video game subtitles. Further details are given in Barbaresi (2014b).

[30] `http://opensubtitles.org`

The subtitle files were searched for using two different sources: sifting through the dumps provided by OpenSubtitles as well as querying the XMLRPC API systematically. The full metadata were also retrieved using the XMLRPC interface for the texts classified as being in German. More details are given p. 204.

4.2.3.3 Between general web crawling and focused retrieval: content discovery of blogs in German under CC license

The problems to be solved in order to be able to build reliable CMC (computer-mediated communication) corpora are closely related to the ones encountered when dealing with general web corpora, described above. The purpose was to design an intelligent crawler targeting specific content types and platforms in order to allow for a fruitful website discovery.

Website discovery First of all, where does one find "German as spoken/written on the web"? Does it even concretely exist or is it rather a continuum? Considering the ongoing shift from web *as* corpus to web *for* corpus, mostly due to an expanding web universe and the potential need for a better text quality, it is obvious that only a small portion of the German web space is to be explored.

Now, it is believed that the plausible distribution of links between hosts follows a power law (Biemann et al., 2013). By way of consequence, one may think of the web graph as a polynuclear structure where the nuclei are quite dense and well-interlinked, with a vast, scattered periphery and probably not so many intermediate pages somewhere in-between. This structure has a tremendous impact on certain crawling strategies. There are ways to analyze these phenomena and to cope with them (Barbaresi, 2014a), the problem is that there are probably different linguistic realities behind link distribution phenomena.[31]

Blog discovery on wordpress.com I chose a specific blogging software, WordPress, and targeted mostly its platform, because this solution compared favorably to other platforms and software in terms of blog number and interoperability. First, `wordpress.com` contains potentially more than 1,350,000 blogs in German[32]. Second, extraction procedures on this website are transferable to a whole range of self-hosted websites using WordPress, allowing to reach various blogger profiles thanks to a comparable if not identical content structure.

The crawl of the wordpress.com website was prepared by regular visits of a tags homepage (`de.wordpress.com/tags/`) listing tags frequently used in German posts. Then, a crawl of the tag pages (such as `de.wordpress.com/tag/gesellschaft/`) enabled us to collect blog URLs as well as further tags. The whole process was used repeatedly to find a total of 158,719 blogs.

The main advantage of this methodology is that it benefits from the robust architecture of wordpress.com, a leading blog platform, as content- and language-filtering are outsourced, which appears efficient.

The discrepancy between the advertised and actual number of blogs can be explained by the lack of incoming links or tags, by a substantial proportion of closed or restricted access blogs, and finally by the relatively short crawl of `wordpress.com` due to politeness rules used.

[31] While these notions of web science may seem abstract, the centrality and weight of a website could be compared to the difference between the language variant of the public speaker of an organization, and the variants among its basis.

[32] `http://wordpress.com/stats`

Blog discovery in the wild A detection phase is needed to be able to observe bloggers "in the wild" without needing to resort to large-scale crawling. In fact, guessing whether a website uses WordPress by analyzing HTML code is straightforward if nothing was done to hide it, which is almost always the case. However, downloading even a reasonable number of web pages may take a lot of time. That is why other techniques have to be found to address this issue.

The detection process is twofold, the first filter is URL-based whereas the final selection uses HTTP HEAD[33] requests. The *permalinks settings*[34] defines five common URL structures for sites powered by WordPress, as well as a vocabulary to write customized ones. A HEAD request fetches the meta-information written in response headers without downloading the actual content, which makes it much faster, but also more resource-friendly, as fewer than three requests per domain name are sufficient.

Finally, the selection is made using a hard-coded decision tree, and the results are processed using the FLUX-toolchain (Filtering and Language identification for URL Crawling Seeds) (Barbaresi, 2013a, 2013b), which includes obvious spam and non-text document filtering, redirection checks, a collection of host- and markup-based data, HTML code stripping, document validity check, and language identification.

Content under CC-license CC-licenses are increasingly popular public copyright licenses that enable the free distribution of an otherwise copyrighted work. A simple way to look for content under CC-licenses consist of scanning for links to the Creative Commons website[35], which proves to be relatively efficient, and is also used for instance by (Lyding et al., 2014). I obtained similar results, with a very good recall and an acceptable precision around .65.

That said, as a notable characteristic of internet content republishing resides in the severe copyright restrictions and potential penalties, we think that each and every blog scheduled for collection has to be carefully verified, an approach in which I differ from (Lyding et al., 2014).

I will describe the results of the manual evaluation phase below. The results of automatic homepage scans on German blogs hosted by wordpress.com show that blogs including commentaries are rather rare, with 12,7 % of the total (20,181 websites); 0,8 % *at best* under CC license (1,201); and 0,2 % *at best* with comments and under CC license (324).

To allow for blog discovery, large URL lists are needed. They were taken from previous webcrawling projects as well as from pages downloaded from wordpress.com. I obtained the following yields: there were more than 10e8 URLs URLs from the CommonCrawl project[36], of which approximately 1500 blogs were mostly written in German and potentially under CC-license. The German Wikipedia links to more than 10e6 web documents outside of the Wikimedia websites, in which 300 potential targets were detected. In a list of links shared on social networks containing more than 10e3 different *domain names*, about 100 interesting ones were found. Last, there were more than 10e6 different URLs in the pages retrieved from wordpress.com, in which more than 500 potentially interesting blogs were detected.

In terms of yield, these results show that it is much more efficient to target a popular blog platform. Social networks monitoring is also a good option. Both yield understandably much more blog links than general URL lists. Even if large URL lists can compete with specific

[33] http://www.w3.org/Protocols/rfc2616/rfc2616
[34] http://codex.wordpress.org/Using_Permalinks
[35] http://creativecommons.org/licenses/
[36] http://commoncrawl.org

searches with respect to the number of blogs discovered, they are much more costly to process. This finding consolidates my conclusions (Barbaresi, 2014a) concerning the relevance of the starting point of a crawl. In short, long crawls have a competitive edge as regards exhaustiveness, but it comes at a price.

The final list of blogs comprised 2727 candidates for license verification, of which 1218 were hosted on `wordpress.com` (45 %).

Conclusion The example of the blogs illustrates well what prequalification means in a restricted context, since at the end of the operation described above, the actual content still has to be retrieved. However, thanks to prequalification, fitting sources for the intended purpose have been found. The next step in the case of the blogs is to process the list of URLs, i.e crawl the blogs, and retrieve all suitable content, which involves a qualification of the blog posts, i.e. corpus building and quality assessment processes.

4.3 Qualification of web corpus documents and web corpus building

4.3.1 General-purpose corpora

4.3.1.1 Theory: definitions and possible categorization

Although the objectives have priority over the typology and over the definitions themselves, it is still useful to redefine a few notions in order to emphasize the fact that the experimental field and the scientific constructs are non-standard, because of their extreme heterogeneity.

These definitions could be neither exhaustive nor applicable to all cases. They are an attempt to formalize different constraints for text inclusion. Depending on a given corpus usage scenario, not all filtering steps pictured in figure 4.5 are of interest. For a large case study using robust tools the mere identification of web texts may prove sufficient. It could even be too restrictive for non-standard text analysis, whereas the notion of corpus which linguists usually refer to implies going all the way down to coherent texts, a definition which may even seem too loose in this perspective.

It is indeed necessary to use either a broader or weaker definition of the notions below in order to fit to peculiarities of “web texts".

Text The contents of a web page qualify as a text if a range of indicators (mostly based on text statistics) yield values that are above a few basic thresholds such as the number of words per line. Therefore, lists of all kinds (address lists seem to be very common on the Internet) should be discarded in this step, as the purpose is to select what could plausibly be a series of paragraphs (with a paragraph consisting of one full sentence at least).

Well-formed text If a given document seems to exhibit mixed content, e.g. full paragraphs and lists, or lists that are bound to contain sentences, it may be tested for well-formedness.

A well-formed text appears to be structured, it may contain titles or quotes but the overall statistical analysis of its features indicates that its shape is regular (for example because the dispersion of the values of statistical indicators remains close to the mean). As this is a formal

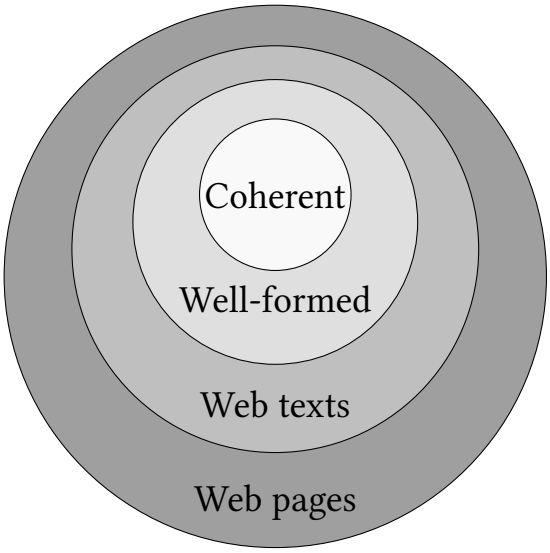


Figure 4.5: Possible selection steps during corpus construction

examination, a decision based on such features may not be fully satisfying, thus requiring finer linguistic cues.

Cohesion and coherence These notions refer to a text that means something to a human reader, which implies that gibberish and nonsensical computer-generated texts should ideally be detected as such and discarded.

Features on discourse level, such as the relative amount of discourse markers and topic analysis, could reveal if a text is satisfying from this point of view. The notion of connected text or connected discourse as it is found in the literature (Marshall & Glock, 1978) can also be productive. However, assessing text cohesion and coherence is a challenging task, which would require extensive work.

Additionally, a web document that appears to contain several parts written in different languages also raises coherence issues.

4.3.1.2 Practice: design decisions in web corpus construction

Introduction It is not always easy to make a decision regarding a text type, let alone a given document. Defining the criteria by which the decision to remove a document is made is also quite difficult. For instance, many documents contain a mix of good and bad segments and thus represent borderline cases. Classification tools, and even human annotators, may fail at assigning a category to cooking recipes for example, as texts of this type are somehow coherent although they rely a lot on lists and can take a lot a different forms, with eminent stylistic differences between enumeration of ingredients and proper advice as to how they are to be prepared.

The decision to systematically remove documents is thus a design decision with major consequences for the composition of the corpus and with potential negative side effects on the distribution of linguistic features. Certain linguistic phenomena might be more or less accidentally underrepresented (for instance concerning the population and/or some specific design criteria) if very long or very short documents are not included, for example. On the

other hand, certain lemmata or parts-of-speech might be overrepresented if long word lists or lists of names are not removed, etc.

The work on design decisions described in this section has been performed in collaboration with Roland Schäfer and Felix Bildhauer at the FU Berlin. It is mostly documented in a joint article, cf (Schäfer et al., 2013).

Approach We defined three particular steps in order to address the issue of design decisions. First, we planned to examine how well humans perform given the task of classifying documents as good or bad web corpus documents. Second, we wanted to introduce and evaluate a completely unsupervised method to classify documents according to a simple but effective metric. Third, we wanted to introduce a format for the representation of corpora in which cleanups like boilerplate detection and text quality assessment are not actually executed as deletion. Instead, we wanted to keep the potentially bad material and mark it as such.

The first and second steps are addressed here.

Inter-annotator agreement protocol Our primary goal in this study was to find out whether corpus designers have clear intuitions about the text quality of web documents, and whether they could operationalize them in a way such that others can reproduce the decisions. Therefore, we randomly selected 1,000 documents from a large breadth-first crawl of the .uk TLD executed with Heritrix. It is the crawl which serves as the basis for our UKCOW2012 and UKCOW2013 corpora (Schäfer & Bildhauer, 2012). The first 500 documents of the sample (in the following to be called “early data") were from the initial phase of the crawl, the second 500 (to be called “late data") from the final phase (after eight days of crawling), when the average quality of the documents is usually much lower (shorter documents, web shops, etc.). The documents were pre-processed for HTML stripping, boilerplate removal, code page normalization, etc., and were thus reduced to plain text with paragraph boundaries.

Then, three coders (A, R, S) were given the task of rating each document on a 5-point scale $[-2..2]$ as to how good a corpus document it is. What we tried to measure was whether the examined documents qualify as corpus texts or not. Coders R and A (respectively Roland Schäfer and myself) were expert annotators, i.e. corpus designers with a shared understanding of what kind of corpus they want to build. Coder S (Sarah Dietzfelbinger) was a student assistant who had previously participated in at least three related but not identical rating tasks on the same kind of data, amounting to at least five work days of coding experience.

A series of criteria was agreed upon, the most important being:

- Documents containing predominantly full sentences are good, “predominantly" meaning considerably more than 50% of the text mass, as perceived by the coder.

- Boilerplate material in sentence form is good, e.g. “You are not allowed to post comments in this forum."
  Other boilerplate material is bad, e.g. “Copyright © 2046 UAC Ltd."

- Sentences truncated or otherwise destroyed by some post-processing method are good as long as they are recognizable as (the rest of) a sentence.

- Repetitions of good sentences are good.

- Decisions should not depend on the length of the document, such that a document containing only one good sentence would still be maximally good.
- Non-English material contributes to badness.
- Non-sentence material (lists, tables, tag clouds) contributes to badness.
- However, if a list etc. is embedded in a coherent text which dominates the document, the document is good (prototypially recipes with a substantial amount of instructions).

The scale is interpreted such that 1 and 2 are assigned to documents which should definitely be included in the corpus, -1 and -2 to documents which should not be included, and 0 to borderline cases. In an initial phase, the coders coded and discussed one hundred documents together (which were not included in the final sample) to make results more consistent.

Coefficient	Early data	Late data
κ (raw)	0.397	0.303
$ICC(C,1)$	0.756	0.679
κ ($t \geq 0$)	0.673	0.625
κ ($t \geq 1$)	0.585	0.555
κ ($t \geq 2$)	0.546	0.354

Table 4.13: Inter-annotator agreement of the three coders for a text quality rating task performed on 1,000 web texts. Below the line are the results for ratings converted to binary decisions, where r $\geq$ n mean that any rating r $\geq$ n was counted as a positive decision; κ is Fleiss' Kappa and ICC the intraclass correlation. (Schäfer, Barbaresi, & Bildhauer, 2013)

Inter-annotator agreement results Table 4.13 summarizes the results. Despite clear guidelines plus the initial training phase, the best value (ICC = 0.756) on the early 500 documents was mediocre. When the documents got worse in general (and also shorter), the confusion rose (ICC = 0.679). Notice also the sharp drop in raw agreement from 0.397 to 0.300 between the early and the late data. Since Fleiss' κ is not very informative on ordinal data and the ICC is rarely reported in the computational linguistics literature, we also converted the coders' ordinal decisions to binary decisions at thresholds of 0, 1, and 2.

The best value was achieved with a threshold of 0, but it is below mediocre: $\kappa = 0.660$ for the whole data set. The value was in fact below the interval suggested in (Krippendorff, 1980) as acceptable. Even if Krippendorff's interval (0.67, 0.8) is not the final (task-independent) word on acceptable κ values as suggested, for example, in (Carletta, 1996) an (Bayerl & Paul, 2011), then 0.660 is still uncomfortably low for the creation of a gold standard. For the binary decisions, the raw agreement also dropped sharply from 0.900 to 0.762 between the early and the late material.

It should be noted that coders judged most documents to be quite acceptable. At a threshold ≥ 0 on the 5-point scale, coder A considered 78.4% good, coder R 73.8%, and coder S 84.9%. Still, there was an 11.1% difference between R and S. Positive decisions by R were almost a perfect subset of those by S, however. In total, 73.0% were rated as good by both coders.

Summary and conclusions All in all, despite written guidelines and training on a test set, we only achieved a rather poor inter-annotator agreement. There was a clear drop in raw agreement between “early" and “late" data, meaning that classifying the content becomes more difficult as the quality drops.

One of the crucial results of this experiment is that corpus designers themselves disagree substantially. Surely, it would be possible to modify and clarify the guidelines, do more training, etc. This would most likely result in higher inter-coder agreement, but it would mean that we operationalize a difficult design decision in one specific way. It has been shown for similar tasks like boilerplate classification that higher inter-coder agreement is possible (Steger & Stemle, 2009). If, however, paragraphs and documents are deleted from the corpus, the users have to agree with the corpus designers on the operationalization of the relevant decisions, or they have to look for different corpora. Our approach (Schäfer et al., 2013) is an attempt to remedy this situation.

Once again, while this study raises mostly technical questions which corpus designers have to care about, I am convinced that linguists working with web corpora should also be aware of how such matters have been dealt with.

Last, the fact that such a classification operation is not a trivial task implies that it is probably difficult to train a classifier which approximates all the decisions. Attempts in this direction feature the search for a common denominator which may be found manually or automatically during the training process, and which may cut through the diverging human decisions in order to find a sort of happy medium.

4.3.1.3 Examples

The examples below are extracted from a crawl of the co.uk top-level domain, a sample of which has been annotated in order to build a corpus construction benchmark described in Schäfer et al. (2013). As I generally choose to discard given text types and consider others as acceptable, this subsection is an attempt to build a typology of these different types. However, one should bear in mind that these decisions may not be consistent with those taken by other researchers. As my colleagues and I showed in Schäfer et al. (2013), the acceptability of web texts in corpora is a surprisingly dividing issue where it is difficult to reach a consensus, even with predefined guidelines (see above).

The practical typology mostly follows the theoretical one (see above). It first distinguishes web documents which are not texts, second texts which are not well-formed, third well-formed texts which are not coherent, and last a supplementary category exhibiting mixed cases which I would describe as undecidable or at least particularly difficult.

Web documents that are not texts The first case of fragments which do not qualify as texts in a linguistic sense are directories, which may be seen as a type of list. In the example below, the directory even include non-standard or faulty tokens, which makes the case even clearer.

> The Corner House Dental And Tooth Whitening Centre
> cosmetic, crown, tooth, caring, crowns, practitioner, smile, surgeon, bleaching, whitening, surbiton, oral, mouth, gums, paste, toothbrush, floss, practise, tolworth, gentle, painless, preventetive

It is hard to tell whether the list has been automatically generated, and for what purpose. It may be a list of keywords destined to complement search engine optimization techniques,

or it may be a list of links. At this stage, i.e. after boilerplate removal and cleaning, this is not clear. The actual text is practically of no linguistic value, even if from the point of view of computational linguistics the faulty tokens might be interesting to test morphological analysis tools or spell-checkers.

Texts that are not well-formed The second example is a little more complicated. If simple criteria, such as the length of paragraphs for instance, are taken into account, it may qualify as a text. However, as the content is actually a list of addresses, it is not of much linguistic value either, a problem which I target using the term well-formedness (see above for a definition of well-formedness).

> Publishers of the periodicals listed:
> Animal Conservation, Zoological Society of London, Regent's Park, London NW1 4RY, U.K.
> De Harpij, Stichting De Harpij, Van Aerssenlaan 49, 3039 KE Rotterdam, The Netherlands.
> International Zoo Yearbook, The Zoological Society of London, Regents Park, London NW1 4RY, U.K.
> Milu, Tierpark Berlin-Friedrichsfelde, Am Tierpark 125, D-1136 Berlin, Germany.
> Oryx, Blackwell Scientific Publications Ltd (for Fauna and Flora Preservation Society), Osney Mead, Oxford OX2 0EL, U.K.
> Der Zoologische Garten, Gustav Fischer Verlag Jena GmbH, Villengang 2, D-07745 Jena, Germany.

In fact, an address list combined with bibliographical references has its interests, and once again in an applied perspective the addresses of different countries offer a very interesting tokenization problem. However, to have tokens such as "NW1" and "Aerssenlaan" would most probably rather raise suspicion about the linguistic resource for most corpus linguists. Although it is virtually impossible to completely avoid including addresses in the final corpus, leaving out such clear cases limits the number of "out-of-vocabulary" tokens in the final product.

Well-formed texts that are not coherent The third example is particularly interesting as it illustrates even more intertwined issues. Not only are classified ads a genre which is difficult to address, the example also shows how boilerplate removal problems affect further decision processes and lastly text quality in the corpus.

Classified ads are frequent on the Internet, and it is hard to tell whether they should be systematically included or not, because this decision mostly grounds on the actual result: are the descriptions detailed? Is there much text? Is the text in the corpus conform to the original or at least acceptable enough?

In that sense, the performance of the processing chain has its importance, and the final result is an aggregate of original writing style and idiolect, content display on the ad platform, as well as markup removal, boilerplate removal, tokenization, and so on.

> Select a country
> Please select a country
> Use smart filters
> Choose a vehicle that meets your expectations
> You can share your results via e-mail or Facebook
> how much is it? x from a standard bt landline, calling an 0844 number will cost you 5p (+vat) per minute at all times. overview vehicle details features/specification about us showroom...

> See description
> Vehicle description
> vehicle print out £3,595.00 renault clio 2.0 16v renaultsport 182, both cup packs,now sold , always looking for new stock, clio 2.0 172,182,197 scroll over the thumbnails to enlarge model year: 2005 mileage: 68,000 miles transmission: manual engine size (in ccm): 1,998 power: 182 bhp fuel: petrol interested? call us 08446638070 ? or 0786 7782189 or email us visit us for a test drive jj automotive view by appointment northolt middlesex ub5 5nw find out where we are 2005 54 reg 182 full service history, cambelt change at 55,506 05/2010 new dephaser+service+ 12 mot 8/2012 stainless steel powerflow exhaust [not loud like some replacement exhausts] service invoice's and old mots a/c climate all working unmarked cup alloys standard clio 182 sport, must be seen px welcme hpi clear credit/debit cards taken warranty facilitys contact john on 07867782189 manufacturer renault model clio type standard car doors 3 number of seats 5 colour black steering wheel right-hand drive mileage 68,000
> There was an error sending the message, please try again later.

In that particular case, though it is difficult to decide, I would advocate against retaining this text in a corpus. Not because the original text is of no interest, but because as far as I am concerned I have the impression that something went wrong during boilerplate removal. On the one hand, the platform makes use of snippets and thus truncates the text at the source ("overview vehicle details features/specification about us showroom..."), and on the other hand the text does not seem to be coherent, there seem to be parts missing.

It would be impossible to discuss each text thoroughly similarly to the argumentation above if the final corpus is to be of a web scale. To implement such complex decision processes in an automatic classifier would also probably be too cumbersome.

Mixed cases In both mixed cases below, explicit decisions have to be taken with an eye on project management. Since the cases are not clear enough to be tackled efficiently, a substantial fraction of web documents is bound to fall under rules applying to these extracts. Moreover, their particular genre is typical for web texts as well as problematic for large-scale linguistic studies.

On the one hand, in the first example it is hard to assess the coherence of the paragraph, partly due to boilerplate removal problems. A potentially large text part is missing:

> Research in the group is supported by the following organizations:
> University economics expert appointed to government advisory panel

The second case illustrates the abundance of copyright notices (and legal texts in general), which are frequent in web documents, sometimes at the bottom of each page:

> TES Editorial © 2012 TSL Education Ltd. All pages of the Website are subject to our terms and conditions and privacy policy. You must not reproduce, duplicate, copy, sell, resell or exploit any material on the Website for any commercial purposes. TSL Education Ltd Registered in England (No 02017289) at 26 Red Lion Square, London, WC1R 4HQ

The main problem with the second case is that it entails perfectly sound text. However, if it is frequently published this way, numerous duplicates will find their way into the corpus, which may distort for instance lexical frequencies.

Another problem with the content is more subtle: the license agreement obviously restricts potential uses of the text, and thus of the corpus. It is not clear whether potential corpus users or the publishing of the corpus will fall under the "commercial purposes" clause.

All in all, the difficulties detailed here show that a web corpus carries with it a sum of decisions which have to be made along its building process, be it in an automatic fashion, by way of classification processes trained and applied for example, or at the level of project management, with crucial decisions concerning the types of content, the notion of text, and potential reuses of the final product.

4.3.1.4 Intrinsic quality assessment

Motivation Text quality evaluation is a major issue in web corpus construction as content quality varies greatly not only between retrieved pages but also within a document. In fact, the notion of "text" itself is problematic when it comes to describe tag clouds, name lists or classified ads for example, as well as all possible scraps and shreds of web content which form the output of crawler and processing tools (cf p. 190 for commented examples). All in all, it is a challenge to define what characteristics a text should have in order to be considered worthy of inclusion in a linguistic corpus.

Basic text characteristics Basic, easily computable characteristics can be helpful in order to detect problems in the corpus and adjust processing accordingly. As shown on p. 2.1, document length can reveal the presence of duplicates. It can be noted that visualization of information otherwise present as columns of figures allows for a first glance at the corpus characteristics. That way, striking phenomena can be spotted at an early stage.

Figures 4.6 and 4.7 illustrate the length distribution across a corpus of web texts in several languages, after basic filtering, and based on information delivered by FLUX. On the left side, a particular segment of the distribution with respect to the number of characters is emphasized, whereas on the right the whole distribution of length is shown on word level. Both series of figures are part of the metadata which FLUX associates with a visited web page. As for the figures on p. 83, they have been extracted out of a full web document collection by grouping the output of FLUX during multiple experimental results on multiple websites and target languages.

The most striking fact is the obvious existence of a relatively high number of duplicates, despite basic deduplication steps on host and document levels (see p. 4.2.1.1). Other major characteristics include the existence of a long tail and the fact that most web pages do not contain many words once they are stripped of the HTML markup.

The mere page length can already be a first step in order to improve a corpus. First of all, it is an indicator to be used in the search for perfect duplicates. Since it is highly improbable that many documents share the exact same length (before, after markup removal, or both), it should rather be an improvement than a degradation to remove part or all of the potential duplicates.

Content hashing, i.e. an algorithmic function with a reproducible output, can also be used to find perfect duplicates based on the bare text (Baroni et al. (2009) delete all the documents detected by this process).

Additionally, the distribution of document lengths can be used to manually or automatically define an adequate window in order to set the desired maximum and minimum length.

Thresholds based on frequent words: Summary of known methods Baroni and Ueyama (2006) and Baroni et al. (2009) both use a list of frequent function words in order to discriminate between types of web texts. The rationale of this approach is that texts using few or no function

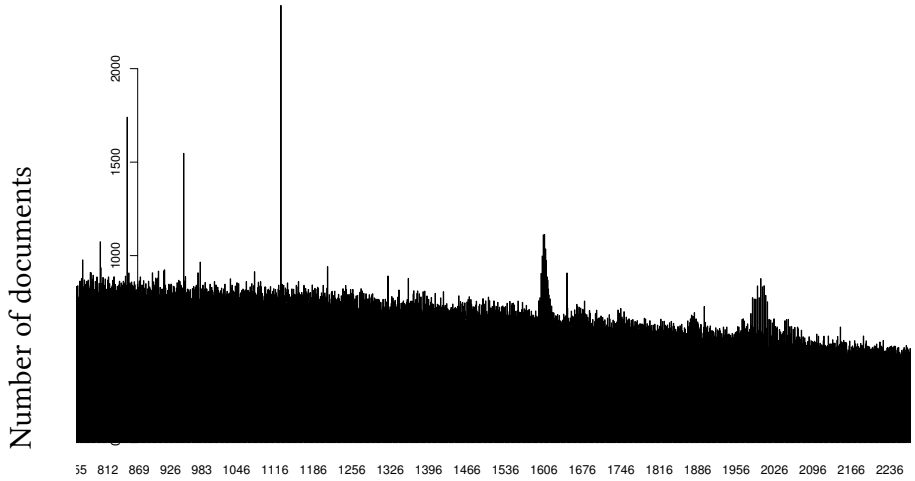


Figure 4.6: Length in characters before markup removal for a particular length segment (zoomed view)

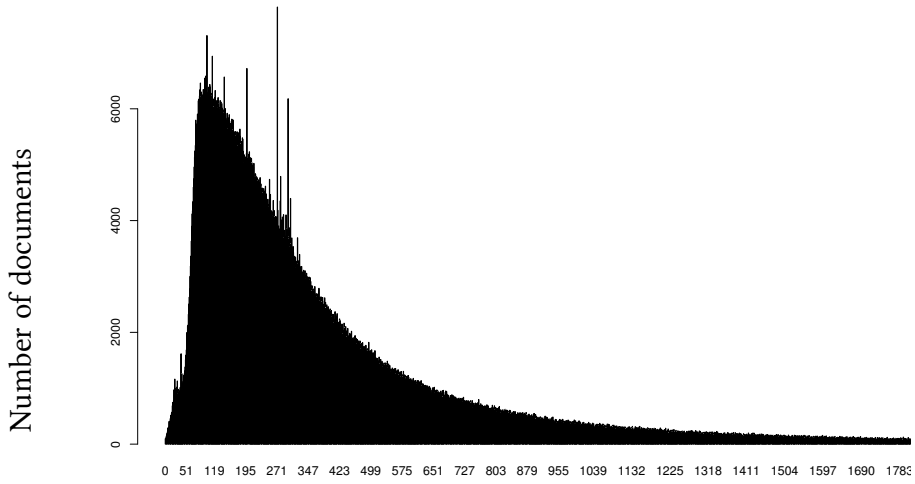


Figure 4.7: Length in tokens after HTML stripping

words are not likely to be interesting from a linguistic point of view. Thus, these texts should be identified and discarded during a specific processing step. The words as well as their desired frequency are computed from previously existing reference corpora.

Schäfer et al. (2013) also use a list-based approach relying on frequent word types. The existence of a single criterion combined with an unsupervised approach allows for a time-efficient implementation with low processing costs. The criterion described in the following is of statistical nature, and as such it is adaptable to different conditions.

In the training phase, the *n* most frequent word types are calculated based on a sample of documents from the corpus. For each of these types, the weighted mean of its relative frequency in the sampled documents and the corresponding weighted standard deviation are calculated (weighted by the length of the document) as an estimate of the corpus mean and standard deviation. In the production run, these two statistical values are used to calculate the normalized deviation of the relative frequency of these *n* types in each corpus document. The more the frequency in the document deviates negatively from the estimated population mean, the worse the document is assumed to be. If the added normalized negative deviation of the *n* types reaches a threshold, the document is removed from the corpus. Both in the training and the production run, documents are processed after markup stripping and boilerplate removal. As expected, both the mean and the standard deviation are relatively stable after 1,000 documents.

In short, the lack of short and otherwise highly frequent words can be measured easily and with consistent results for the kind of data in which one is interested. The tricky part of the approach is a matter of threshold: finding the most interesting and/or most productive value for the filter cannot be determined *in abstracto*. Experiments are necessary in order to manually assess the outcome of this particular filtering phase.

4.3.1.5 Improving intrinsic quality assessment and decision processes

In continuation of the work performed with Roland Schäfer and Felix Bildhauer on web text filtering (Schäfer et al., 2013), I have tried to improve the way texts are selected. In fact, thanks to the information we gathered on web corpora, it was clear that a considerable amount of text is filtered out, so that it should be worth removing it after the computational effort download and preprocessing represent. Thanks to the manual evaluation used in the same article on English web texts of the *.co.uk* TLD, a calibrated testing ground is available, which can be used to benchmark different approaches with respect to their efficiency.

I will now introduce and evaluate a number of easily extractable document features which help to improve automatic classification of web texts with respect to a gold standard annotated by experts.

Approach Like Schäfer et al. (2013), I adopted an unsupervised approach to document quality assessment and went under the similar assumption that the lack of short and otherwise highly frequent words can be measured with consistent results. The difference in my approach resided in the introduction of several variables, as a single criterion can be seen as a strength and as a weakness at the same time. My analysis was also motivated by the evaluation of linguistic features which have a certain relevance at discourse level. Such features are used for instance in order to assess the readability (or processing difficulty) of texts.

The main approaches of readability studies were summed up in a previous chapter. Not only is the methodology interesting as an example of current research in NLP, it is also relevant for accessing useful text characteristics, so that part of the methods described by researchers have been used in this study, in the context of web text qualification.

Similarly to the readability score for example, the output of a web text filter is basically an estimation of several parameters that lead to an approximation as to how relevant a given web text can be for corpus building.

There are several similarities between my approach and the method used in Schäfer et al. (2013). First, it was meant to be an unsupervised approach to document quality assessment.

Second, it relied on the assumption that the lack of short and otherwise highly frequent words can be measured with consistent results, so that the frequent words can be used as a way to grasp the nature of an unknown text. This is also a common feature with Baroni et al. (2009).

The main differences were twofold. First, I decided to introduce several variables, which also meant that a balance of several indicators had to be achieved. Second, I tried to find linguistic features which may have a certain relevance at discourse level. The second point could be of particular importance, since manual examination of the dataset indicated that a decisive criterion seemed to be text cohesion (see above).

Potential indicators tested The following features were tested for statistical significance against the Schäfer et al. (2013) data set consisting of cleaned web page contents, most notably using variable selection methods (Hastie, Tibshirani, & Friedman, 2009):

General indicators: text length, number of tokens, number of punctuation marks per token, proportion of capital letters among all characters, proportion of spaces.

Indicators inspired by readability studies: number of tokens per line (average and standard deviation)[37], sentence length (average and standard deviation).

Discourse-related indicators: proportion of discourse markers as defined by (Fraser, 1999), proportion of temporal markers.

Unigram-based indicators: most experts agree that the well-known nonsense formula "etaoin shrdlu" captures the most frequent characters in English (Salomon & Motta, 2010, p. 4), so that ten characters were tested: *e*, *t*, *a*, *o*, *i*, *n*, *s*, *h*, *r*, and *l*.

Discourse markers The discourse-related indicators as defined by Fraser (1999) included 34 different words which could be considered in at least one of their senses to be discourse related. However, due to the obvious polysemy of most short words, a study on token-level may not capture the various uses of discourse markers. Additionally, the category of discourse markers is difficult to define precisely, meaning that the list is not necessarily exhaustive, nor a perfect fit for this category. I took the article by Fraser (1999) as a reference, as it is an acknowledged work on this topic.

Here are the tokens which were counted relatively to the total number of tokens in order to build a discourse-related indicator: accordingly, also, although, analogously, and, as, because, besides, but, consequently, conversely, equally, hence, however, likewise, moreover, namely, nevertheless, nonetheless, otherwise, similarly, rather, since, so, still, then, therefore, though, thus, too, well, whereas, yet.

Temporal markers Concerning the temporal markers, I did not find any authoritative study on the topic, i.e. a list which I could take for granted. There are lists of this kind, but none of them seemed complete or acknowledged enough among linguists to be taken as a reference. Thus, I decided to compile a list by myself, using several word lists for students learning English as a foreign language.

[37]Usually, there are line-breaks in the original HTML. There are also potential line-breaks in the markup, e.g. $< br/ >$ or $< p >$ tags, which are translated to newlines in the conversion from HTML to text.

As a matter of fact, this list is not perfect either, there are cases, such as “finally", where the word could as well be understood as a discourse marker, and there is no evidence whether these borderline cases are more frequent in a sense or another. Similarly, “yet" is present in the discourse list but not in this one, although the word can be used in a temporal sense.

The final list of temporal markers comprises 28 different tokens: after, afterward, afterwards, always, before, beforehand, directly, earlier, early, finally, immediately, in-between, later, meanwhile, never, nowadays, occasionally, previously, sometimes, soon, sooner, suddenly, today, tomorrow, tonight, while, whilst, yesterday.

Concerning both lists, the items are invariable and do not comprise any spaces. Thus, the difference between word and token is not relevant in this case, that is why I use both terms without distinction in this particular context.

Even if the lists are not indisputable, they give a general idea, since most selected terms lean in a certain direction, be it as a discourse or as a temporal marker.

Feature selection The indicators mentioned above do have something to do with text structure and fabric, yet they are not necessarily good descriptors of the characteristics relevant to decision processes in web corpus construction. Because web texts include a wide variety of texts, including new genres and potentially unforeseen text types, it would be hard to determine *a priori* which indicators are to be used. That is why experiments have to be made in order to assess the impact of these potential indicators and to discriminate among them.

Several methods have been used in order to select the most relevant indicators. Most of them are of a statistical nature. In order to check for obvious predictions and/or redundancy, I first drew correlation matrices for all observed indicators. Correlation implies a statistical relationship between variables, but not necessarily a direct or even indirect causation. Thus, the values in the matrix are to be interpreted with care.

In order to be able to assess the impact of the variables, visualization offers a practical solution, as the human eye usually identifies patterns more quickly in a graphic representation of a table than in a table containing figures. The advantages of visualization in corpus linguistics are discussed later in the remainder of this document.

For example, figure 4.8 shows the statistical relationships between the proportions of all letters of the alphabet in the test set, i.e. the manually annotated corpus of 1,000 web pages in English. Last, the fact that a document is being kept in the final corpus or rejected is also examined, it is expressed by the so-called “choice" variable, which is 1 if the mean score of the annotators was superior to zero, and 0 otherwise.

The white color denotes a weak positive correlation, while the light blue color denotes a strong positive correlation. Pink stands for a weak to moderate negative correlation. Each variable obtains a correlation coefficient of 1.0 with itself, that is why the diagonal is blue. In the other cases, the different colors show that the distribution of characters in the corpus is not random. For instance, where the mouse pointer is, i.e. at the intersection between h and u, one may say that dependence between h and v does not seem to be of any interest, while the dependence is noticeable between h and t, and h and u, with positive and negative coefficients respectively.

However, the interesting part of figure 4.8 concerns the choice variable: although there are no distinctively strong characters which could be used directly as predictors, the variation in the correlation ratios shows that a few characters may be more useful than others when it comes to predicting the outcome of the classification.

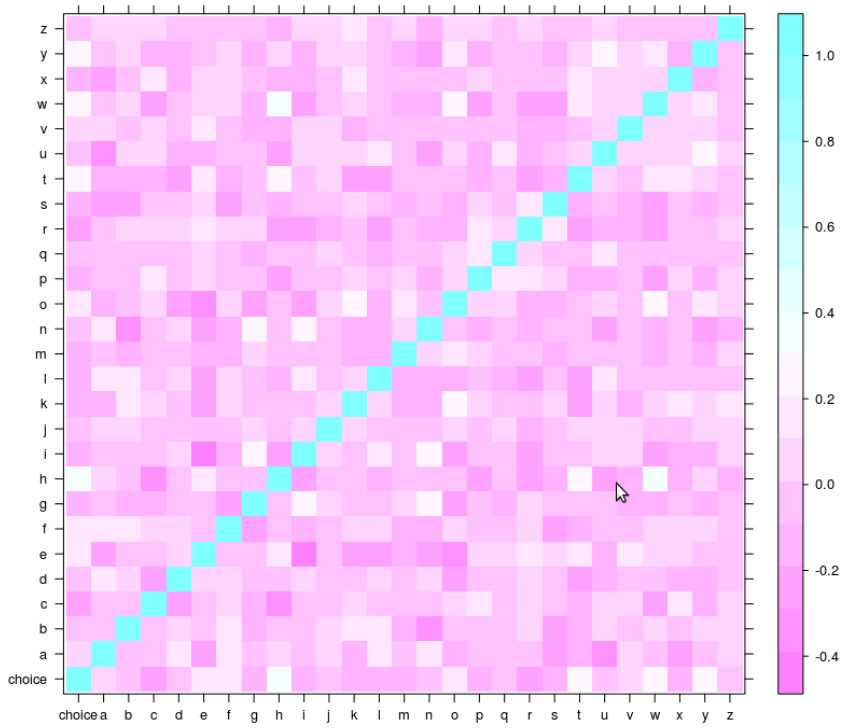


Figure 4.8: Correlation matrix for characters from a to z and choice variable

In a way, the examination of statistical relationships between characters and outcome validates the “frequent characters" criterion, but it also shows that said criterion should not be given much importance.

In figure 4.9, the colors illustrate the statistical relationships between a selection of variables. More precisely, a positive correlation is depicted in red and a negative one in dark blue. The score is not a binary choice this time as in the example above, but the mean of the three assessors’ scores.

For reasons of clarity, variable names had to be shortened, *t-len* stands for text length in characters, *n-words* stands for the number of words (or more precisely tokens). *w/line* stands for the mean number of words per line, while *w/line-sdev* denotes the standard deviation. *disc-mark* and *temp-mark* denote respectively the discourse and temporal markers defined above. *punct-%* stands for the amount of punctuation (as defined by the POSIX class). *w/sent* and *w/sent-sdev* stand for the mean and standard deviation of tokens per sentence, although the sentence boundary detection applied here is basic.

First, one notices obvious correlations such as the length in characters and the number of words per text. But similarly to figure 4.9, the actual interest of the graph is in looking at the values correlating with the score. The values are promising since the range of coefficients is higher than for the characters of the alphabet. While this observation does not necessarily imply that there is a direct relationship, certain indicators may be more relevant than others.

The amount of punctuation for instance is negatively correlated with the score, making up the strongest negative correlation of the table. On the other hand, while means and standard deviations both offer potentially interesting ratios, the mean seems to be superior in both cases in this respect. Concerning the discourse and temporal markers, the first seem to be more of interest than the latter.

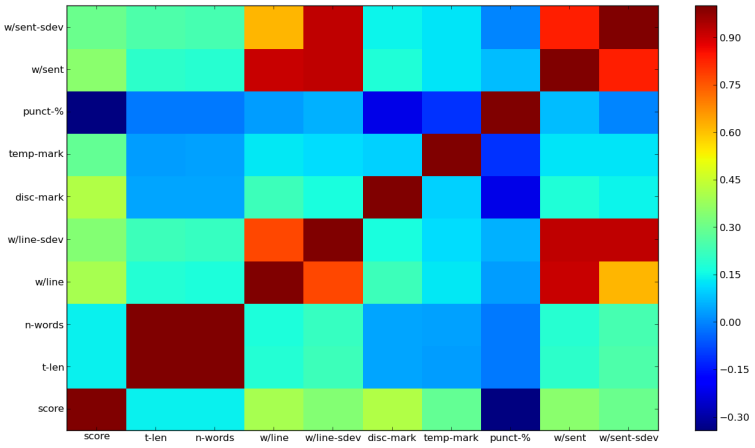


Figure 4.9: Correlation matrix for selected indicators

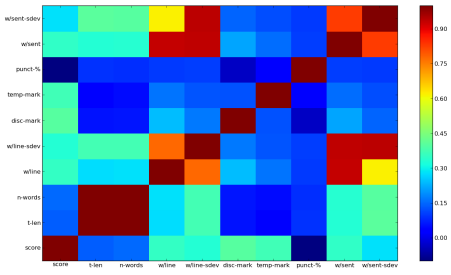


Figure 4.10: log1, first subset of 500 web documents

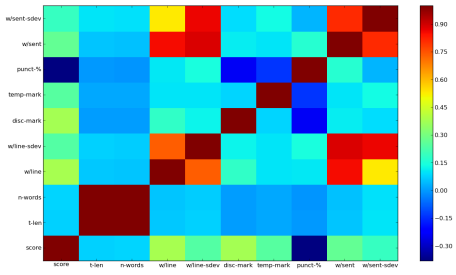


Figure 4.11: log2, second subset of 500 web documents

The figures 4.10 and 4.11 show the same results as figure 4.8, but for each subcorpus. In fact, figure 4.8 is actually an aggregate, since the document collection is made up of 500 documents sampled from the beginning of the crawl and 500 documents sampled from the end. Experience has shown that the quality is much worse at the end of the crawl (see above), but also that the nature of documents changes.

In that case, it is interesting to see if the indicators capture the changes between the beginning and the end of a crawl. Indeed, it seems to be the case, even if the colors are not to be trusted completely since the scale changes. There are negative correlations in figure 4.11, while it does not seem to be the case in figure 4.10. The conclusions of the aggregate graph seem to be valid in this case too: the means are slightly superior to the standard deviations as an indicator, and so are the discourse markers relatively to the temporal markers.

Now that a first impression has shown that the selected indicators could be significant, the relevant variables for this purpose have to be selected. More importantly, since the purpose is to use several indicators at once in order to determine if a text is to be discarded or kept, the selection should also find the right combination of indicators.

In order to do so, I use a statistical variable selection based on regression analysis, which is

a statistical process for estimating the relationships among variables. It is much more powerful than the manual assessment of a correlation matrix since this type of statistical model allows assessment of the statistical significance of variables, their explanatory power compared to the others.

More precisely, the type of regression used here is logistic regression, a probabilistic statistical classification model (Hastie et al., 2009). Logistic regression measures the relationship between a categorical dependent variable and one or more independent variables. The dependent variable here is binary. Models based on logistic regression are used throughout various disciplines, the most salient example in NLP are probably conditional random fields (CRF), which apply logistic regression on sequences. In NLP, the sequences can be composed of words and/or tags, for instance in the case of part-of-speech tagging. CRFs have gained popularity since the end of the 2000s.

After testing several kinds of models and comparing their efficiency, I decided to use a generalized linear model. Compared to other available alternatives, this type of model cuts a plane, which is what the classification operation by machine learning techniques is generally about, all the more since the classification to be made is binary: inside or outside the document collection.

A linear regression for instance provided a reasonable estimate for the text collection but there was room for improvement. Since a logistic function generally cuts the plane by drawing a curve, it seems a better match for the properties of the collection. Moreover, approaches based on statistical significance were preferred to machine learning techniques since they do not follow a black box model. While most machine learning algorithms do not provide feedback on the adjustments they make in order to provide a better fit for the data, statistical classification provides a number of metrics in order to assess the importance of the variables and the resulting explanatory power of the whole model. Moreover, it is possible to better assess the general efficiency of the model by performing cross-validation tests, as reported below.

In order to assess the impact of the variables automatically, the stepping process as implemented in R (R Core Team, 2012a) has been used. This algorithm selects the most relevant variables using basic statistical tests. While feature selection is generally considered as a complex process in NLP, this simple approach is strictly applicative and was chosen for its simplicity: using the statistical indicators at hand, the stepping algorithm is about finding a minimal subset of features while leaving the explanatory power of the model nearly untouched. It can be seen as a way to remove statistical impurities by focusing on the features providing the clearest signal.

Final combination The final combination is a heterogeneous blend to be used on document level.

- Indicators computed on the whole text
  - Amount of punctuation with respect to the total amount of tokens
  - Amount of spaces with respect to the total amount of tokens
- Line-based indicator (since there are potential line-breaks in the original HTML as well as in the markup)
  - Mean number of tokens per line

- Token-based indicator
  - Proportion of discourse markers per token
- Character-based indicators
  - Proportion of upper case characters with respect to the total number of alphabetic characters
  - Proportion of *t* with respect to the total number of alphabetic characters

Since the model is not a black box, or at least not completely obscure with respect to its functioning, it is possible to comment on the relevancy of the selected indicators. It can be said that all indicators proved to be highly relevant save for the proportion of *t*, which is perceived by the model as being useful but significantly less important than the others. Manual tests proved that contrary to other characters of the alphabet the proportion of *t* really yielded better results, that is why it has been used in the final selection.

As seen in the correlation matrices, the discourse markers proved to be better at explaining the score than the temporal markers. In the end, the first are even a highly reliable indicator, while the latter are not strictly necessary for the model to function and were thus set aside.

It is conceivable that the clearest signals come from the variables whose measurement is less disputable. The manually compiled list of temporal markers proved inferior to the acknowledged list of discourse markers.

Similarly, the basic sentence boundary detection did not prove as efficient as the line-based metric.

Coefficients	Estimate	Std. Error	z value	Pr($>\|z\|$)	Signif.
(Intercept)	-10.83718	1.81442	-5.973	2.33e-09	***
Words per line	0.06103	0.01222	4.993	5.94e-07	***
Discourse markers	0.21692	0.06510	3.332	0.000861	***
Punctuation	-0.04472	0.01206	-3.707	0.000210	***
Spaces	0.73631	0.10446	7.048	1.81e-12	***
Upper case characters	-0.35689	0.05547	-6.434	1.24e-10	***
Proportion of Ts	0.19862	0.09725	2.042	0.041120	*

Table 4.14: Results of logistic regression analysis for selected features and choices made by rater S (the “naive" coder).
Significance codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Figure 4.14 shows the results of a regression analysis, similar to others in corpus linguistics literature, e.g. (Bader & Häussler, 2010). According to the statistical indicators, all selected features can be expected to have a significant impact on the results.

The proportion of Ts is not as important as the rest, however, similarly to the other variables, as removing it causes a drop in performance.

Evaluation In comparison with the method described in the original experiment on decision processes (Schäfer et al., 2013), results detailed in table 4.15 show a slightly lower recall but a better precision and accuracy, so that the F-score is better overall.

The homogenous results among the coders are satisfying, since the inter-annotator agreement is rather low (see above). The fact that both F-score and actual accuracy have been improved show that it is imaginable that the model may have found a common denominator for all three coders, which would not have been visible to the naked eye.

In fact, cross-validation is a technique used to determine how a given model will perform in practice, on data other than that used for training. During the so-called 10-fold cross-validation, ten different subsamples are created, and one single subsample is used to evaluate the model. This is also known as the "leave one out" technique, and the rest is used to generate the model. The score at the end of the procedure is an average of all folds.

For this case, the cross-validation results are roughly comparable to the accuracy, with differences as low as 0.001, 0.004, and 0.001 for A, R, and S respectively. These results indicate that the model is bound to generalize well.

	Prec.	*Diff.*	**Rec.**	*Diff.*	**F1**	*Diff.*	**Acc.**	*Diff.*	**10-fold CV**
A	.963	+ .049	.921	− .038	.941	+ .005	.906	+ .018	0.905
R	.955	+ .099	.904	− .069	.929	+ .018	.892	+ .041	0.888
S	.951	+ .143	.971	− .005	.961	+ .077	.934	+ .123	0.933

Table 4.15: Evaluation of the model and comparison with the results of Schäfer et al. (2013). A, R and S stand for three different coders. *Prec* denotes the precision, *Rec* the recall, *F1* the F-score, and *CV* the cross-validation based on accuracy scores.

Similar results around or above 90% accuracy were obtained using decision tree learning with the same criteria, more precisely recursive partitioning and regression trees (see figure 4.12 for an illustration of regression trees). However, in that case no cross-validation tests are possible, and overfitting cannot be excluded. Nonetheless, these results show that the better performance of the model is not an artifact of the chosen classification method.

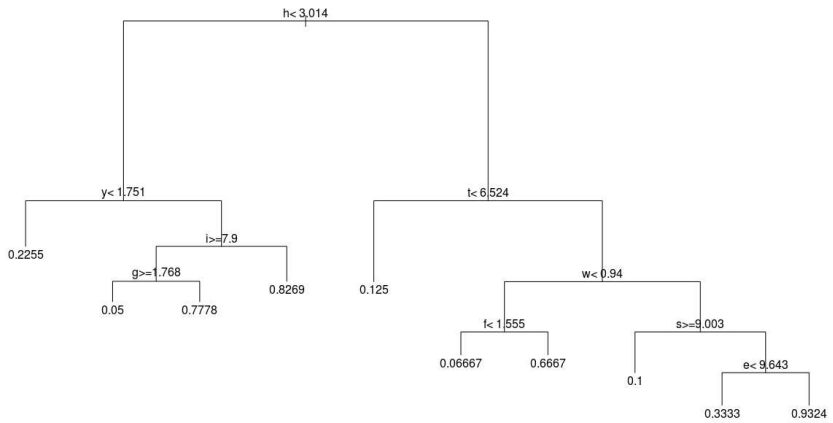


Figure 4.12: Example of decision tree for the characters etaoin shrdlu

As expected, discourse markers seem to be good predictors of text quality, mostly because they may help to sort out texts featuring no cohesion at all. The lack of relevance of temporal markers might be explained by the text type, as they could be used to discriminate between narrative text and spam for example, which was not the case here.

The proportion of *t* accounted for about 2% more accuracy, the proportion of *r* and *d* also had a statistical significance. Consequently, future work may include bigram and trigram analyses, already used for language identification purposes and which may also be useful in this case.

Conclusion The model introduced in this section provides a better basis for design decisions. Following the model, decision processes can be improved, since a possible common denominator was found between several diverging annotation policies. The most positive result is that despite the low inter-annotator agreement, the F-score improved for all raters with only a minimal drop in recall for two raters, which may indicate that this model is close to the common denominator of the raters' decisions.

Indicators constructed on elementary text and discourse analysis can be relevant. All in all, easily extractable variables which can be determined with precision are in demand. The main advantages of this method are twofold.

First, it remains fast as the indicators are surface features which can be extracted easily, and as such it is applicable to big data processing.

Second, another possible advantage is that most indicators could be used across languages, the only ones needing adaptation would be the character-based ones. Terminology extraction for European languages using the Wiktionary[38] has been attempted and implemented in FLUX: once the language has been identified, corresponding discourse and temporal markers are used to compute the ratios used in the model.

It has not been tested whether the translations of markers yield viable indicators. Their inclusion in FLUX, along with other indicators, is part of a general policy consisting in computing indicators and encoding them as metadata for each web page that is being analyzed. Then, potential users are left with options concerning metadata usage in decision processes or text qualification in general, as well as the eventual weights given to chosen indicators.

At the end, a few open questions concerning design decisions remain. First, could several steps be useful to include or exclude texts? The single step described here does not reflect the theoretical partition described above, it compresses the four characteristics needed in order to be part of a selective corpus into one single evaluation step.

Second, is it possible and desirable to filter at paragraph or sentence level? My approach is to filter texts on text-level, since I wish to favor text integrity. However, in more traditional corpora, it is not rare to find excerpts or even scraps of original texts. In that case, more selective processes are needed, since the number of potential candidates increases exponentially.

Third, should we corpus designers filter aggressively or should we try to maximize the recall? This is a delicate issue which closely depends on potential uses of web corpora. More opportunistic approaches probably need as much text as possible, regardless of minor irregularities in quality, while others may need texts loser to the existing written standard.

4.3.2 Specialized corpora

4.3.2.1 Dealing with non-standard and unclean text: the example of a subtitle corpus

Overview of corpus building process Corpus building included the following main phases, which are detailed in the sections below:

[38] `https://en.wiktionary.org`

1. Search for subtitles files
2. Retrieval of metadata
3. Data download
4. Data processing

The processing chain takes text files as input, more precisely several formats of subtitles (MicroDVD, SubViewer, SAMI, SSA, TXT). The output is in the form of text files (TXT format) or XML format following the TEI guidelines.[39]

Retrieval The subtitles are retrieved from the OpenSubtitles project[40], a community-based web platform for the distribution of movie and video game subtitles.

The subtitle files are searched for using two different sources: first by sifting through the dumps provided by OpenSubtitles and carrying out cross checks to discover other resources; secondly by querying the XMLRPC API systematically, i.e. for each known subtitle ID, in order to find those in German, according to the metadata.

The full metadata are also retrieved using the XMLRPC interface for the texts classified as being in German. Each video document is identified by an IMDB number which could theoretically make metadata completion using other sources possible (for example `imdb.com` itself).

The drawbacks discovered during retrieval are twofold: on the one hand there are growing restrictions on download frequency, and on the other hand the quality of the website in terms of information architecture could be improved (database access is sometimes inconsistent).

Processing steps Data processing encompasses the following major steps:

1. Normalization
    - Unicode conversion and repair
      The default working format is UTF-8.
    - Identification of subtitle format
      There are five main known formats: MicroDVD, SubViewer, SAMI, SSA, and TXT.
2. Text cleaning
    - Removal of markup and text cleaning, based on file format detection
      Mostly time specifications, advertisements, typography, and so-called ASCII art.
3. Formatting of the output
    - Optional fusion of frames into sentences
      Performed by a basic sentence boundary detection.
    - Optional conversion from text to XML TEI format.

[39] `http://www.tei-c.org/`
[40] `http://opensubtitles.org`

Example

Raw data:	**Result:**
909 01:28:16,334 --> 01:28:19,202 Ich genieße einfach den Rest des Sommers. 910 01:36:09,932 --> 01:36:13,141 Copyright EUROTAPE 2013 Untertitel: Cosima Ertl u. a.	Ich genieße einfach den Rest des Sommers.

Known issues and design decisions Format-related issues included the existence of several formats as well as encoding and markup irregularities in both encoding and markup, so that robustness was paramount.

There are obviously cases where UNIX-tools such as *file* and *iconv* fail to detect the proper encoding or to translate it properly, probably because of previous incorrectly assessed encodings. A typical case for instance are files which were most probably natively encoded in Windows-/CP-1252 but then processed and destructively re-encoded as being latin-1/ISO-8859-1. LACLOS does not fix this problem, it merely contains the damage by applying a series of oneliners.

Content-related issues include the existence of several sets of subtitles for the same film requiring heuristics to choose the potentially better one, flaws of OCR methods used on subtitles which require error correction, multilingual documents, and spam or advertising.

- Partially addressed issues

  1. There may be several versions (i.e. files) for the same film, although this problem rarely occurs for German subtitles since they are ten times less frequent than for example English ones. In the cases where several files are available, heuristics can be used to choose between the different versions. The default is to select the subtitle file most often downloadeded in the past.
  2. Multi-lingual documents (see quality assessment below)
  3. Spam or advertising, most often for a subtitle “brand" or a movie release team, including exotic markup, which is easy to detect, but also for full sentences such as “*Normalerweise hat Qualität ihren Preis ... / doch bei uns kriegt ihr sie umsonst !*"[41]

- Problems left untouched

  1. Some files are the result of an optical character recognition which failed partially, leaving vowels out or turning all “i" into “l". There was no attempt to remedy this.

[41] “Normally, quality does have a price tag... / but not with us!"

2. Cases have been reported where the DVD-menu and not the actual content of the subtitles have been framed. The files which are concerned are easy to filter out based on their size. There is no automatic procedure to see if a better subtitle file is available and/or to replace it.

No normalization of any kind was attempted on token level, which means that possibly divergent orthographic forms, be it because of linguistic variants or because of typos or digitalization mistakes, are left as such.

The paratext within subtitles consists of scene descriptions and indications for the hearing impaired. It is not a clear case of markup, since it is linked to film and subtitle content and usually written in plain English.

Nonetheless, as it does not correspond to actual utterances in the video, this paratext was excluded from the content for psycholinguistic reasons and marked as such in the XML export version of the corpus. This was done everywhere the paratext was clearly identifiable, for instance because of particular punctuation styles, but not in other cases, which included for example story introduction at the beginning of a film or epilogues at the end.

4.3.2.2 Constitution of a republishable corpus: the example of blogs

Problems to solve The problems to be solved in order to be able to build reliable computer-mediated communication (CMC) corpora are closely related to those described above, encountered when dealing with general web corpora. Specific issues are threefold. First, what is relevant content and where is it to be found? Second, how can information extraction issues be tackled? Last, is it possible to get a reasonable image of the result in terms of text quality and diversity?

Approach I will present three possible ways to cope with the issues described above. First, I will design an intelligent crawler targeting specific content types and platforms in order to allow for productive website discovery (see p. 184) and, second, to allow for the crafting of special crawling and content extraction tools. Third, I will find metrics to compare Internet-based resources with already known, established corpora, and assess their suitability for linguistic studies (see p. 220).

Manual assessment of content and licenses Due to the necessity of having an error-free classification of republishable content, the actual presence of CC licenses on pages detected as such was verified manually. More specifically, this was the occasion for starting a blog classification to be performed manually using a series of predefined criteria dealing with (1) general classification, (2) content description, and (3) determination of authorship.

First, concerning the general classification, the essential criteria were whether there is really something to see on the page (e.g. no tests such as *lorem ipsum*) and whether it is really a blog. Another classification factor is whether the blog has been created or modified recently (i.e. after 2010-01-01).

Second, concerning the content description, the sine qua nons are to check that the page contains texts, the majority of which needed to be in German, and that the text content is under a CC license. Other points were whether the webpage appeared to be spam, whether the content could be clearly classified as dealing with Germany, Switzerland or Austria, whether

the content appeared to be *Hochdeutsch* or a particular dialect/sociolect, and last if the website targeted a particular age group such as kids or young adults.

Third, the authorship criteria were twofold: was the blog a product of paid, professional editing or did it appear to be a hobby; and was the author identifiably a woman, a man or a collective?

Concerning the essential criteria, the results of the classification were such that 1,766 blogs could be used without restriction (65%), since all the textual content qualified for archiving, meaning that there was indeed text on the webpage, that it was a blog (it contained posts), that it was mostly written in German and under CC license.

BY-NC-SA	652
BY-NC-ND	532
BY-SA	351
BY	282
BY-NC	129
BY-ND	58

Table 4.16: Most frequent license types

DE	1497
Unknown	715
AT	146
CH	69
LU	2
NL	2

Table 4.17: Most frequent countries (ISO code)

The breakdown of license types is shown in table 4.16, as are the results of country classification in table 4.17.

The CC licensing can be considered to be a sure fact, since the CC license can theoretically not be overridden once the content has been published. Possible differences between adaptations of the license in the various countries should not be an issue either, because it is done in a quite homogeneous way. The relatively high proportion of BY-NC-ND licenses (30%) is remarkable. While the "-ND" (no derivative works) restriction does not hinder republication as such, its compatibility with corpus building and annotation is unclear, so that such texts ought to be treated with caution.

Examples of blogs For screenshots of the following blogs see appendices (see p. 244):

- agchemludwigshafen.wordpress.com
- behindespace.wordpress.com
- bilderplan.wordpress.com
- blog.beetlebum.de

- blog.foxxnet.de
- blog.nihonnikonni.com
- derfuhs.wordpress.com
- dmhdf.wordpress.com
- reizgesteuert.wordpress.com
- saitam.wordpress.com
- sunnyromy.wordpress.com

Examples of typical utterances Among the top morphological analysis failures (posts)

- 's wär' nunmal sooo soooo soooooo
- $>:D$ $:-($
- HIV-positiver
- Rüebli Röteli
- Ubuntu Myspace Wikileaks xkcd Shitstorm

Among the top morphological analysis failures (comments)

- m.E. evtl. ggf. bspw wg zb vll/vllt usw
- kannste immernoch brauchts fürn aufm
- *g* <3 ;-(
- jaaaa
- Rezäpt
- WinXP

Conclusion The results show that it is possible to find blogs in German under Creative Commons license. The crawling and extraction tools seem to give a reasonable image of blog language, despite the fact that the CC license restriction impedes exploration in partly unknown ways and probably induces sociological biases.

4.3.2.3 Intrinsic quality assessment

Indicators used for quality assessment In general, user-generated content on the Web comes with an inherent unevenness to be smoothed out. Subtitles are no exception, they can be of different origin and nature, but also mixed quality, so that design decisions are not necessarily clear-cut.

Most of the indicators are token-based, they can be roughly split into the following categories:

- N-gram analysis (from tokens/unigrams to 5-gram tokens)
- Language identification (spell checker and probabilistic models)
- Annotation toolchain and analysis of results (elementary text statistics)

A significant proportion of unknown words as well as the presence of words in a concurrent language, in that case English, are indicators that can automatically trigger the exclusion of texts above a certain threshold. Additionally, the output of the language identification system `langid.py` (Lui & Baldwin, 2012), i.e. a language code and a confidence interval, is used to find outliers in the subtitles collection.

About 10% of the original files are not used because of encoding errors, improper OCR-use but mostly because they were detected as not being in German.

N-gram analysis The first and most efficient way to go about it is probably the analysis of the frequent n-grams over the whole corpus, i.e. all the supposedly clean subtitles. Major encoding or content irregularities are easily detected in a unigram list. Others are more subtle and require skimming through the bottom of the list to explore the less frequent 3-, 4- or 5-grams. This is an important step as such corpora may be used to build word frequency profiles, and from this point of view a noisy subtitle corpus is worthless.

This screening takes place after a basic tokenization process, more elaborate options are not needed as this step reveals major failures in corpus construction.

Language identification Another way to look for potential errors is to use a spell checker like *hunspell*. If a high proportion of tokens in a given text of the corpus are marked as errors, the text may be not be in the target language. However, this measure is clearly destructive, as the interest of subtitle corpora resides precisely in their unexpected “new" or “fresh" content, be it words or multi-word expressions, which are unknown to processing tools.

Nonetheless, spell checkers are faster and often sufficient: if the output of the German spell checker is that 20% of the tokens are unknown while the output of the English spell checker marks 70% as such, then no further analysis is required. The same happens if both checkers failed to recognize more than 80% of the text.

Thus, while this tool often leads to relevant hints, a spell checker could not be used in an unsupervised approach. It is advisable to check the results manually for each text that is allegedly out of norm.

Language identification software is also a relevant approach, but it has the same disadvantages as the spellcheckers. The method used is often probabilistic, which is why the confidence

interval is highly relevant. While spell checkers can be used where the possible concurrent languages (in this case English) are known, a proper language identification might be necessary to identify texts in a wide array of languages.

Annotation toolchain and text statistics In order to be considered fully processed, the corpus has to satisfy formal constraints such as clean encoding and format. Results of the processing chain can be used for qualitative evaluation (for example minor errors), but also quantitative evaluation, as they provide synthetic indicators on a range of text characteristics.

Among the available indicators, the following have been regularly used:

- Mean word and sentence length (including punctuation)
- Distribution of part-of-speech tags (in percent): NN/NE (common nouns and named entities), verbal forms, punctuation, function words
- Measure of lexical diversity: for several sample sizes, calculate a type/token ratio in order to compare different type of resources
- Percent of tokens for which the morphological analysis failed

4.4 Corpus exploitation: typology, analysis, and visualization

4.4.1 A few specific corpora constructed during this thesis

Synoptic tables of corpora I have worked upon and of tools I have worked with are to be found below in the appendices (see p. 238).

4.4.1.1 Press

Methodology I will start with a well-known (because it is traditional) type of (web) corpus: newspaper corpora (see appendices for a list of available corpora).

A few technicalities are common to most newspaper corpora gathered on the Web:

Precision & recall Precision is usually preferred to recall, which means that getting complete articles is considered more important than gathering all articles without distinction.

Deduplication A hash function is used to shorten the links and make sure a given URL is retrieved just once. Nonetheless, there is still duplicate content for various reasons, including lack of overview on the publisher side and refactoring of articles under different titles. Titles, excerpts (if the case applies) and text hashes altogether can be used to deduplicate the texts in the corpus. Further analysis may show that there are very similar texts in the corpus, which may be explained by genre-specific features.

Boilerplate removal The uninteresting (boilerplate) parts of a web page are cut off using specially crafted regular expressions, which makes the process close to web scraping techniques.

Setting The crawlers are relatively fast (even if they were not set up for speed) and do not need a lot of computational resources. They may be run on a personal computer.

Although a newspaper corpus cannot be republished as is, parts of it can be made available upon request, for instance in the form of scrambled sentences or as derivative material such as n-gram lists. For this very reason, the crawl of newspaper websites should not be too easy to spot, nor should it consist of an exact copy of the whole website.

In fact, getting an exact copy is difficult, since preparing a specialized crawl implies to reverse-engineer the way a given content management system deals with document publishing. It may be impossible to understand completely the logic of a given website and to discover all pages, even if no internal links point to them. However, sometimes the attentive recipient of the web page gets to know the taxonomy of a website very well as well as its link structure, so that orphans or articles neglected by the publishers themselves are found.

On the other hand, there may also be superfluous texts in the linguists' collection. Since downloading and indexing the content takes time, the end corpus does not reflect the exact content of a newspaper's website. Pages may be taken offline because of legal reasons, or renamed, while previously unknown parts of an archive may be put online.

Two comparable corpora of German newspaper text gathered on the web: Bild & Die Zeit
I used two comparable corpora of German newspaper text, built from on the daily tabloid *Bild* and the weekly newspaper *Die Zeit* respectively.

Two specialized crawlers and corpus builders were designed in order to crawl the domain names bild.de and zeit.de with the object of gathering as many complete articles as possible. A high content quality was made possible by the specially designed boilerplate removal and metadata recording code.

At the time of writing, the last version for *Bild* was from 2011 and the last version for *Die Zeit* was from early 2013. The corpora feature a total of respectively 60,476 and 134,222 articles.

Whereas the crawler designed for *Bild* has been discontinued due to frequent layout changes on the website, the other one concerning *Die Zeit* is still actively maintained, its code has been made available under an open source license.[42]

Choice of the newspapers The primary goal was the construction of a reference corpus of comparable newspaper text. Both newspapers were chosen accordingly, since they account for two specific strategies in the general press.

There are striking differences in the writing styles of the articles. In *Bild* there is a tendency towards very short sentences and the use of colons and illustrated content (multimedia components were not part of the crawling strategy). Comparatively, there is a lot more textual content in *Die Zeit* with a proclivity towards subordination and elaborate sentences as well as long and detailed articles. In short, it seems that *Bild* uses a lot more parataxis whereas there is a lot more hypotaxis to be found in *Die Zeit*.

The expected audiences of these newspapers are also quite different. While both can be considered successful (with *Bild* being one of the most popular newspapers in Europe, selling about five times more copies than *Die Zeit*), the first one aims to keep its leading status of popular tabloid, while the latter is well-regarded for its journalistic quality.

[42] https://code.google.com/p/zeitcrawler

Last, there are also technical reasons explaining this choice: on both sides, crawling was and is not explicitly forbidden, which is not always the case with German newspaper websites. As a matter of fact, no hindrance was encountered while performing the crawls.

Statistics *Bild* corpus (2011 version):
There are 60 476 unique documents in the corpus, which were published between November 2007 and November 2010 according to the metadata. The corpus comprises 19 404 260 tokens.

Die Zeit corpus (early 2013 version):
There are 134 222 unique documents in the corpus, which were published between 1946 and January 2013 according to the metadata, with the major part of the articles being published later than 2005. The corpus comprises 105 456 462 tokens.

Access The *Bild* corpus has been used internally at the ENS Lyon but has not been made accessible publicly due, on the one hand, to a lack of adequate infrastructure, and on the other hand to the existence of similar corpora in German linguistic research.

The *Zeit* corpus has been adapted to suit the needs of a larger user base and an advanced corpus-querying infrastructure. Finally, it has evolved over the course of time as new articles were added. It is presently possible to query the corpus via the search portal of the DWDS project.[43]

4.4.1.2 German Political Speeches Corpus

Interest To my knowledge, no corpus of this kind has so far been made publicly available for German. There are corpora containing political campaign speeches partly developed by commercial companies as well as different sources that gather political texts classified as important, but not systematically with a common reference.

Another main interest of this corpus is that most speeches could not be found on Google until I put them online, and the Chancellery speeches from before 2011 are not to be found on its website anymore.

Last, there is no copyright on this corpus, which is quite rare for German texts. As they were given in public, all the speeches can be freely republished as stated by German copyright law[44]. Nonetheless, the law indicates that a republication must not target a particular author.

As a matter of fact, the whole corpus has become part of a reference corpus for German, DeReKo, compiled at the IDS Mannheim.[45]

XML format The corpus was released in XML and Unicode format. There is one XML file grouping all the texts of each subcorpus together, since I thought it was easier to manipulate that way (its size stays reasonable). The files have their own DTD, inspired by the TEI guidelines. The corpus is not fully TEI-compliant yet, but it is closer than the first release.

Raw XML file The metadata are properly encoded, the texts are given as they were crawled, with no enrichment whatsoever.

[43] http://www.dwds.de

[44] § 48 UrhG, Öffentliche Reden.
http://bundesrecht.juris.de/urhg/__48.html

[45] http://www1.ids-mannheim.de/kl/projekte/korpora.html

XML file Tokenization[46], POS-Tags and Lemmata[47] are included.

4.4.1.3 Subtitles

Interest In the context of psychological research, text corpora such as subtitles are used to derive frequency lists. The frequencies are used to test correlations, with reaction and latency times gained in experiments based on lexical decision and/or naming. In that sense, the quality of the corpus resource is related to its explanatory power and predictive potential.

More specifically, Brysbaert and New (2009) showed word frequencies gained from movie subtitles were superior to frequencies from classical sources in explaining variance in the analysis of reaction times from lexical decision experiments. In fact, the explanatory power of subtitles in psychological experiments has been found to be high, not only for American English, used in the first experiment (Brysbaert & New, 2009), but for many languages, ranging from Dutch (Keuleers, Brysbaert, & New, 2010) to Chinese (Cai & Brysbaert, 2010).

The reason for this superiority is still somewhat unclear (Brysbaert et al., 2011). It may stem from the fact that subtitles resemble spoken language, while traditional corpora are mainly compiled from written language (Heister & Kliegl, 2012). In that sense, it seems feasible to draw an analogy between subtitles and spoken language.

Besides, subtitle corpora may also be relevant to linguistic studies, not only in the form from frequency lists. They may offer a more down-to-earth language sample closer to everyday life and spoken corpora. This is at least what the psychological studies suggest, saying that the subtitles are better predictors because they are less abstract than traditional written corpora.

Potential advantages for lexicography include the discovery of new words and senses, or example sentences for words/senses which are known to exist but cannot be found in standard written corpora.

A general linguistic interest resides in the investigation of language use beyond traditional written corpora, as well as in the exploration of language patterns close to or derived from spoken variants.

Potential interests in computational linguistics include language modeling, since there are tasks for which subtitles corpora may perform better than other types of corpora, and tools hardening, concerning morphology or word sense disambiguation for instance.

At the BBAW[48], the construction of a subtitles corpus originated with the DLexDB project (Heister et al., 2011). Its purpose is to complement the use of the DWDS corpus (Geyken, 2007) to derive frequency lists. Moreover, the subtitles have found their way into comparative studies concerning the corpora of the DWDS project (Barbaresi, 2014a) and other specific web corpora ("A one-pass valency-oriented chunker for German", n.d.).

Concrete expectations Among the expectations linked to the subtitle corpus was the establishment of a clean reference for experiments using psycholinguistic data. Additionally, with this particular register, there is the wish to grasp another language sample which gives a new view on language, as it is supposed to be more realistic, and closer to everyday reality.

[46]Partly using a Perl script developed by Stefanie Dipper, which can be found on her website : http://www.linguistics.ruhr-uni-bochum.de/ dipper/

[47]Provided by the TreeTagger : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[48]Berlin-Brandenburg Academy of Sciences

The corpus is supposed to be clean and homogeneous enough for inclusion in reference corpora to enable research on lexicography. To this end, a version in XML TEI format is available, whereas token and lemma frequency profiles were calculated using a raw text version.

Last, by using metadata corresponding to the subtitles and matching it to other resources, we expected to extract a subset of subtitle files which correspond to films for children.

Result (status: October 2013) 11,956 files were downloaded. According to `http://www.opensubtitles.org/de/statistics` there were 17,116 available subtitles for German at that point. The missing files were not to be found, neither in the dumps nor using the API.

Since most of the indicators for internal quality assessment are token-based, they can be roughly split into the following categories: N-gram analysis (from tokens/unigrams to 5-grams of tokens), language identification (spell checker and probabilistic models), and annotation toolchain and analysis of results (elementary text statistics).

10,795 documents remained at the end of the processing chain, meaning that a total of 1,161 files (9.7%) were blacklisted because of encoding errors, improper OCR-use but mostly because they were detected as not being in German (see indicators below).

The corpus size was 56,276,568 tokens, which makes it an interesting resource, since there are probably enough different texts and enough tokens to cover various enunciation situations as well as to provide somewhat reliable word frequencies[49], provided the text quality is high enough. In the next section I will describe on what basis the quality was assessed.

Error analysis indicated that around 1.5% of the texts still contained irregularities such as markup or encoding errors. More precisely, several thousands tokens were concerned, representing 0.001% of the collection. As these errors should reach thresholds of statistical significance during corpus use, and given the difficulties described above and unclean raw data, the corpus was considered to be exploitable.

No data source was found to solve the data problem, since the data available on the reference platform IMDB cannot be freely used. Web scraping was theoretically an option, but would mean infringing the terms of service.

I did not find any source fulfilling all the necessary criteria, i.e. a good coverage, automatable querying process, and freely available data. Thus, all metadata was extracted from OpenSubtitles, even if in most cases it lacked information and/or did not offer proper genre classification.

Software The software used to download and preprocess the subtitles, LACLOS[50], is available under an open source license: `https://github.com/adbar/laclos`

Results of quality assessment The n-gram analysis is generally useful to exclude ads of all kinds as well as material linked to this specific text genre.

In the case of subtitles, the sentences are noticeably shorter than in other corpora, making it a hallmark of this register (see statistics in comparison of corpora p. 221). Moreover, the type-token ratio is rather low. These results correspond to the assumption that subtitles use less elaborate words and expressions and that they contain a subset of the target language which is supposed to be close to everyday language.

[49] Brysbaert et al. expect such a limit to be around 10-15 MTokens.

[50] The acronym stands for LAnguage-CLassified OpenSubtitles.

There was nothing noteworthy concerning the POS-tags distribution. It was in line with other corpora except for the more frequent occurrence of punctuation, which is a direct consequence of the shorter sentences. The values for noun and verbal forms as well as those for the function words were on par with other corpora. This can be interpreted as an indication that the corpus is regular and clean with respect to a certain language standard.

The morphological analysis failed slightly more often than in the reference corpus (p. 221). The two most likely reasons for this phenomenon were the subtitles containing more abbreviations than other corpora on the one hand, and on the other hand the fact that this linguistic variant is supposed to be closer to a foreign language, including new vocabulary (regarding technology for instance), foreign words as well as colloquial expressions. All of them are not bound to be part of the standard cases used to build the morphological analyzer.

4.4.2 Interfaces and exploitation

4.4.2.1 Introduction

Most of the time, the boundaries between corpus components and/or corpora are drawn manually, starting from a text genre for instance. In that case, the notion of genre is believed to be stable and established enough to describe particular properties of the texts themselves. Texts of a certain genre are expected to form a coherent set.

However, criteria based on expectations towards text types are of an extrinsic nature and do not necessarily reflect actual text characteristics. More precisely, it is unknown whether the texts in such a collection are tied together by common features which can be revealed by quantitative indicators.

There might in fact be remarkable differences within the same corpus. By contrast, there might not be a strict separation between several types of corpora, although they are expected to be different.

The point of this particular subsection is to explore possible ways to find similarities or differences between the texts in a corpus, by developing and using interfaces to access the corpora. First, keywords are computed and used to provide an interaction with a corpus. Second, quantitative indicators are compared and represented in a graphical fashion in order to compare several specialized web corpora relatively to each other.

4.4.2.2 Keyword selection and visualization

First, a series of keywords were selected in order to explore the German Political Speeches Corpus. The purpose was to verify the quality of text and metadata, to give an insight into corpus contents, and to provide an access to corpus data which does not strictly operate from a linguistic perspective only.

The keywords were manually picked from a computer-generated keyword list, in order to reflect main themes that were also relevant to a historical point of view.

The first selection was based on the surface parser described in the appendices (see below). Its purpose was to identify phrases by using the output of a part-of-speech tagger. As a result, salient nominal and prepositional phrases in particular were found, making it possible to extract heads of noun phrases. These heads are supposed to carry more weight in the course of the text than other word types or nouns that are not the head of a phrase.

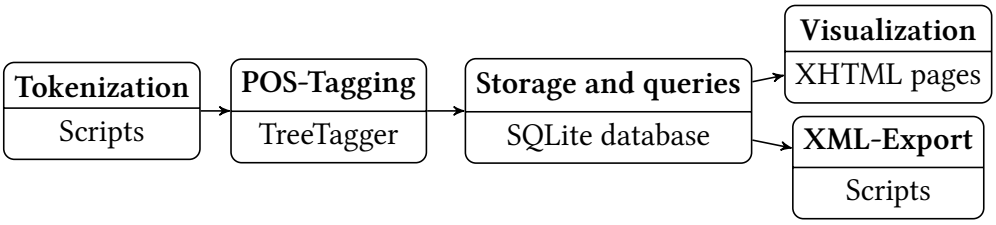


Figure 4.13: Processing steps leading to the German Political Speeches Corpus

The extracted list of frequent lemmata was then used to manually pick a series of words which themselves were assumed to be absolutely central, and which became keywords for the visualization. The rest of the lemmata were used to provide a short description of the texts, as they may give a hint about its content. The five most frequent lemmata were actually used as a part of the metadata corresponding to each text.

I aimed to provide raw data for researchers able to use it and a simple visualization interface for those who want to get a glimpse of what is in the corpus before downloading it or thinking about using more complete tools (cf figure 4.13).

The output is in valid CSS/XHTML format, it uses tabbed navigation and takes advantage of recent standards. It is light both in size and in client-side computation needs, using just a little JavaScript.

The data can be sorted by year, name or text. The word frequency is displayed using histograms. This process makes it easier to look for distinctive and/or relevant keywords. A glimpse of the co-text is also available.

Figure 4.14 shows the list of chosen words in the *Bundespräsidenten* corpus[51].

Description of the interface The second release is the current one, most of the processes have been stabilized but a few tools are still under development.

[51] `http://perso.ens-lyon.fr/adrien.barbaresi/corpora/BP/wortliste.html`

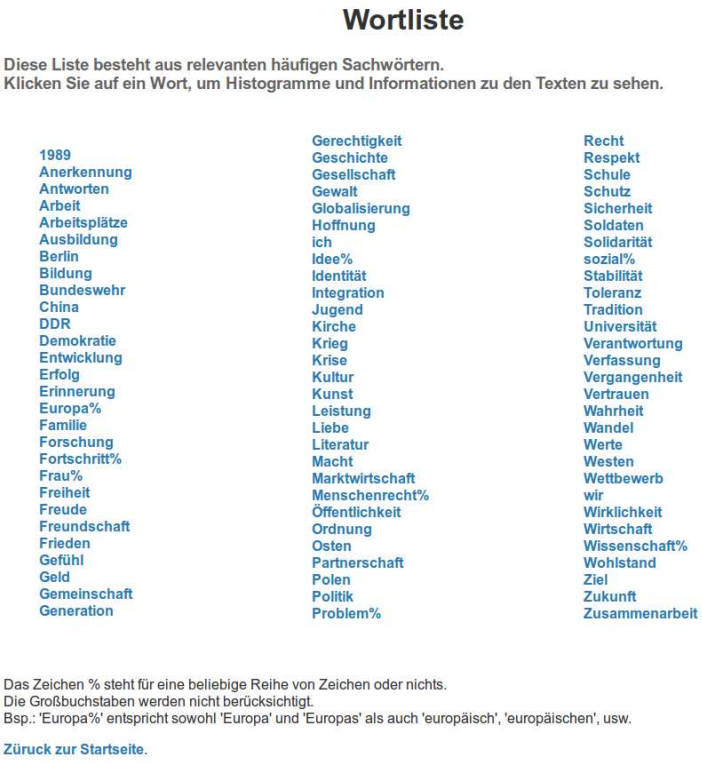


Figure 4.14: List of referenced types for the *Bundespräsidenten* corpus

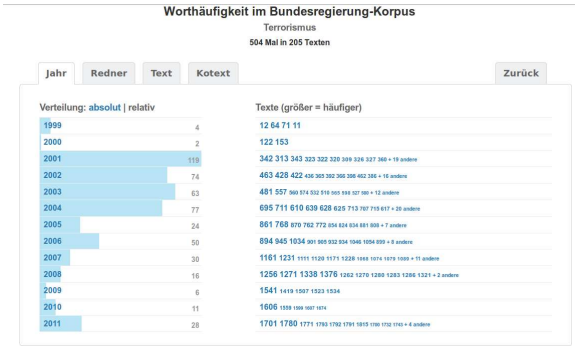


Figure 4.15: First view of the word profile, the bars display relative frequencies

Figure 4.16: View classified by politician, relative frequencies

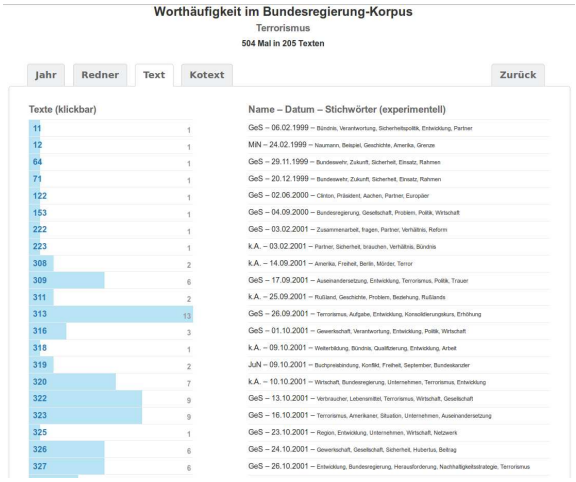


Figure 4.17: Text view with abbreviations of names and keywords

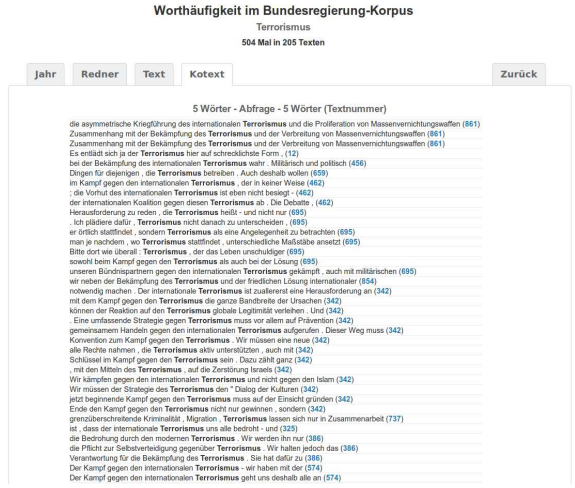


Figure 4.18: Content view displaying left and right context

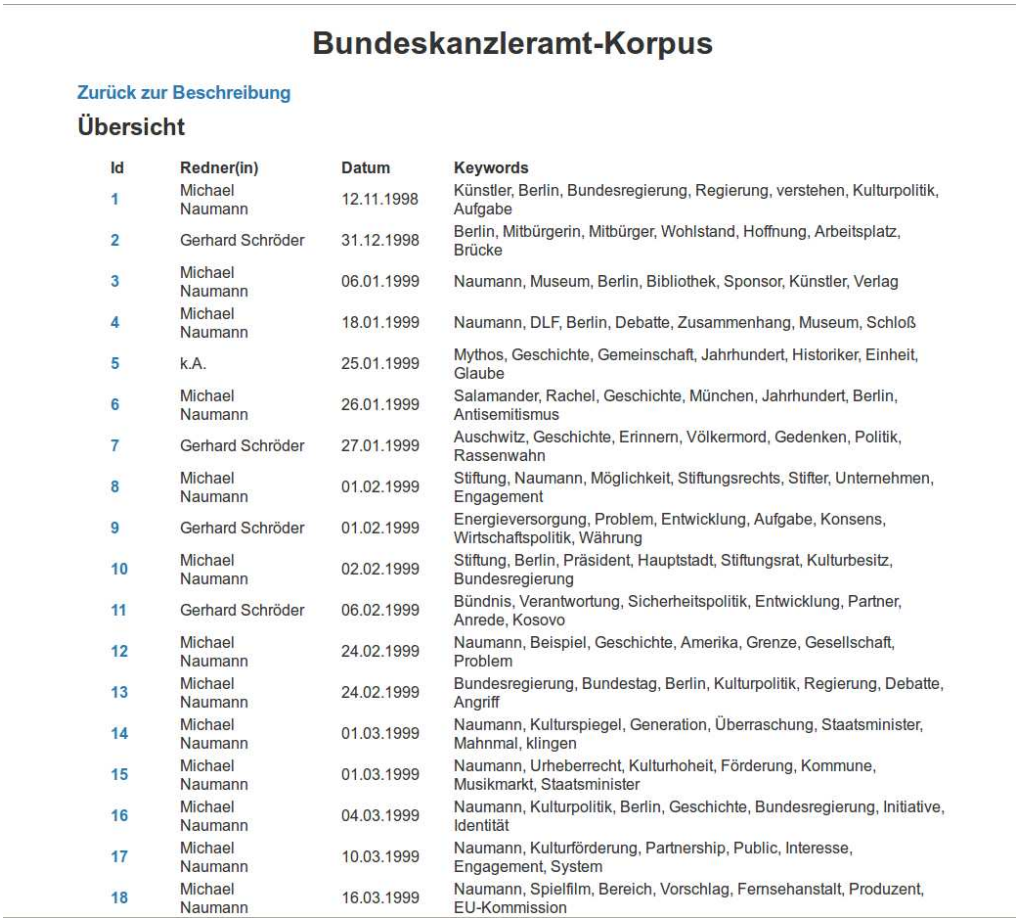
Static web pages are used to distribute the content: a series of queries were performed to generate web pages describing word frequencies. The details on how the web pages are completed on-the-fly using JavaScript to make them lighter and more responsive is described in the appendices below (see below). In fact, JavaScript is used to ensure tabbed navigation, to complete the pages on the fly and to highlight words in the texts.

The interface also provides the user with the context, more specifically the co-text, i.e. five words before, five words after and a link to the text.

Figure 4.15 shows the results for the word *Terrorismus*[52]. A drastic increase in word use in 2001 as well as a subsequent decrease shows that the metadata are probably correct and exemplifies the interest of such a view.

Determination of keywords A list serves as a menu, it contains selected relevant words extracted using a surface parser, which is also described in detail in the appendices (see below).

I designed an algorithm to try and assign relevant keywords to each text. It is based on shallow parsing using the POS-tags. The goal is to look for frequent lexical heads as well as important verbs. I used a stoplist to filter out very common words like “Nation", “Deutschland", “Europa", “Mensch" and verbs like “werden", “können" or “wollen".



Bundeskanzleramt-Korpus

Zurück zur Beschreibung

Übersicht

Id	Redner(in)	Datum	Keywords
1	Michael Naumann	12.11.1998	Künstler, Berlin, Bundesregierung, Regierung, verstehen, Kulturpolitik, Aufgabe
2	Gerhard Schröder	31.12.1998	Berlin, Mitbürgerin, Mitbürger, Wohlstand, Hoffnung, Arbeitsplatz, Brücke
3	Michael Naumann	06.01.1999	Naumann, Museum, Berlin, Bibliothek, Sponsor, Künstler, Verlag
4	Michael Naumann	18.01.1999	Naumann, DLF, Berlin, Debatte, Zusammenhang, Museum, Schloß
5	k.A.	25.01.1999	Mythos, Geschichte, Gemeinschaft, Jahrhundert, Historiker, Einheit, Glaube
6	Michael Naumann	26.01.1999	Salamander, Rachel, Geschichte, München, Jahrhundert, Berlin, Antisemitismus
7	Gerhard Schröder	27.01.1999	Auschwitz, Geschichte, Erinnern, Völkermord, Gedenken, Politik, Rassenwahn
8	Michael Naumann	01.02.1999	Stiftung, Naumann, Möglichkeit, Stiftungsrechts, Stifter, Unternehmen, Engagement
9	Gerhard Schröder	01.02.1999	Energieversorgung, Problem, Entwicklung, Aufgabe, Konsens, Wirtschaftspolitik, Währung
10	Michael Naumann	02.02.1999	Stiftung, Berlin, Präsident, Hauptstadt, Stiftungsrat, Kulturbesitz, Bundesregierung
11	Gerhard Schröder	06.02.1999	Bündnis, Verantwortung, Sicherheitspolitik, Entwicklung, Partner, Anrede, Kosovo
12	Michael Naumann	24.02.1999	Naumann, Beispiel, Geschichte, Amerika, Grenze, Gesellschaft, Problem
13	Michael Naumann	24.02.1999	Bundesregierung, Bundestag, Berlin, Kulturpolitik, Regierung, Debatte, Angriff
14	Michael Naumann	01.03.1999	Naumann, Kulturspiegel, Generation, Überraschung, Staatsminister, Mahnmal, klingen
15	Michael Naumann	01.03.1999	Naumann, Urheberrecht, Kulturhoheit, Förderung, Kommune, Musikmarkt, Staatsminister
16	Michael Naumann	04.03.1999	Naumann, Kulturpolitik, Berlin, Geschichte, Bundesregierung, Initiative, Identität
17	Michael Naumann	10.03.1999	Naumann, Kulturförderung, Partnership, Public, Interesse, Engagement, System
18	Michael Naumann	16.03.1999	Naumann, Spielfilm, Bereich, Vorschlag, Fernsehanstalt, Produzent, EU-Kommission

Figure 4.19: Beginning of the text overview, *Bundesregierung* corpus

[52] http://perso.ens-lyon.fr/adrien.barbaresi/corpora/BR/Terrorismus-rel.html

The first eight words by order of frequency (and relevance) appear in the general overview of the texts, whereas the first five can be found in the representation of the query by texts.

However, it is not only about counting words and determining the most relevant this way. By using syntactic information, the purpose is to give an insight on the topics developed by a government official and on the evolution in the use of general concepts (like security, Europe, freedom or war), which turns it in a way into sort of *Zeitgeist*.

Figure 4.19 shows the beginning of the text overview [53].

4.4.2.3 Extrinsic quality assessment, evaluation, and comparison of corpora

One perfectly legitimate question regarding corpora from the web is whether it is possible to get a reasonable image of the result in terms of text quality and diversity.

Beyond a mere quality competition, corpus benchmarking can have real benefits. First, it can show to what extent web corpora can be used in more traditional corpus studies. Second, it can help to qualify the actual content of a corpus as they are often too big to be examined manually with any desirable precision, for instance by classifying them on the axis of already available resources. Third, it can help uncovering possible flaws in corpus composition or annotation.

Thus, extrinsic quality assessment is mainly about finding metrics to evaluate corpus quality and suitability for linguistic studies.

In order to assess the legitimacy of web corpora, a comparison on lexical level has been made by (Baroni & Ueyama, 2006), for instance:

> "As a sanity check on our procedure, we compared the 30 most frequent words from both corpora." (Baroni & Ueyama, 2006, p. 35)

In that sense, the approach by Baroni and Ueyama (2006) addresses points (1) and (3) above, by checking the corpus for eventual flaws and showing that it is suitable in a corpus linguistics perspective.

Baroni et al. (2009) goes even further in the direction of (1), showing that there may be some work to do to convince linguists. The authors used lexical overlap with reference corpora as an indicator to benchmark corpora from the web, for instance a comparison of a list of nouns with the BNC. These measures based on frequency are used to convince potential users by proving that corpora from the web are not structurally different from more traditional corpora.

Undoubtedly, quality of content extraction has an effect on text quality, since the presence of boilerplate (HTML code and superfluous text) or the absence of significant text segments hinder linguistic work. Moreover, there are intrinsic factors speaking against web texts (see above).

I will show a series of statistical analyses to give the reader an idea of the properties of the crawled corpora. These analyses include comparisons with a German newspaper corpus supposed to represent standard written German.[54]

Materials We presented a series of statistical analyses to get a glimpse of the characteristics of the crawled corpora. Content was divided into two different parts, the blog posts (BP), and

[53] `https://perso.ens-lyon.fr/adrien.barbaresi/corpora/BR/uebersicht.html`

[54] The quantitative analyses and visualizations below are the result of joint work with Kay-Michael Würzner at the BBAW, which is documented in Barbaresi and Würzner (2014)

the blog comments (BC), which do not necessarily share authorship. Due to the relatively slow download of the whole blogs due to crawling politeness settings, we analyzed a subset of 696 blogs hosted on *wordpress.com* and 280 other WordPress blogs. We could not calculate how synchronous the subtitles were with the blogs, manual analysis revealed a high proportion of TV series broadcast in the last few years.

Newspaper corpus The results were compared with established text genres. First, a newspaper corpus is supposed to represent standard written German, extracted from the weekly newspaper *Die ZEIT*, more precisely the *ZEIT online* section (ZO), which features texts dedicated to online publishing (for an example see p. 258). Newspaper articles are easy to date, and we chose to use a subset ranging from 2010 up to and including 2013, which roughly matches both the size and publishing dates of the blogs. They were been digitally generated and are free of detection errors typical for retro-digitized newspaper corpora. ZO is in general considered to be a medium aiming at well-educated people. Therefore, we picked it as a corpus representing standard educated German.

Subtitle corpus Secondly, we used a subtitle corpus (OS) which is believed to offer a more down-to-earth language sample. The subtitles were retrieved from the OpenSubtitles project[55], a community-based web platform for the distribution of movie and video game subtitles. They were then preprocessed and quality controlled (Barbaresi, 2014b). Subtitles as linguistic corpora have gained attention through the work of Brysbaert and colleagues (Brysbaert & New, 2009) who showed word frequencies extracted from movie subtitles to be superior to frequencies from classical sources in explaining variance in the analysis of reaction times from lexical decision experiments. The reason for this superiority is still somewhat unclear (Brysbaert et al., 2011). It may stem from the fact that subtitles resemble spoken language, while traditional corpora are mainly compiled from written language (Heister & Kliegl, 2012). The analogy between subtitles and spoken language was also the primary motivation to include the OpenSubtitles corpus in the following analyses.

The corpora used in this study are all corpora from the Web. Structural properties of the corpora are shown in table 4.18. Their sizes are roughly comparable.

Preprocessing and Annotation All corpora were automatically split into tokens and sentences with the help of WASTE, Word and Sentence Tokenization Estimator (Jurish & Würzner, 2013), a statistical tokenizing approach based on a Hidden Markov Model (HMM), using the standard DTiger model. Subsequently, the resulting tokens were assigned possible PoS tags and corresponding lemmata by the morphological analysis system TAGH (Geyken & Hanneforth, 2006). The HMM tagger *moot* (Jurish, 2003) then selected the most probable PoS tag for each token given its sentential context. In cases of multiple lemmata per best tag we chose the one with the lowest edit distance to the original token's surface.

Analyses All corpora were aggregated on the level of types, lemmata and annotated types (i.e. type-PoS-lemma triplets) resulting in three different frequency mappings per corpus. Analyses were carried out using the statistical computing environment **R** (R Core Team, 2012b).

[55] http://opensubtitles.org

Corpus	Size	mean TL	mean SL	unkn. T
Token level				
BP	33.0	4.95	20.3	2.76
BC	12.8	4.68	16.0†	2.75
ZO	38.2	5.08	17.5	0.89
OS	67.2	3.90	7.6	1.31
Type level				
BP	1.10	11.3	n/a	24.4
BC	0.56	10.5	n/a	27.3
ZO	0.98	12.2	n/a	13.7
OS	0.83	10.1	n/a	23.9

Size ... Number of tokens (resp. types) in the corpus in millions
TL ... Length of token (resp. type) in characters
SL ... Length of sentences in tokens
unkn. T ... Proportion of tokens (resp. types) unknown to TAGH

† Sentence length was re-computed using a statistical tokenization model (Jurish & Würzner, 2013) trained on the Dortmund Chat Corpus (Beißwenger, 2007). The original value using the standard newspaper model was 22.5, a dubious value.

Table 4.18: Various properties of the examined corpora.

Quantitative Corpus Properties Table 4.18 summarizes a number of standard corpus characteristics. Token and type counts as well as length measures including punctuation. While token length is comparable in all four corpora, sentences in the subtitles are less than half as long as in the other corpora. The proportion of unknown types with respect to the standard-oriented morphological analyzer TAGH is by far smaller in the ZEIT corpus and marginally higher in blog comments than in the other standard-deviating corpora.

Type-Token Ratio Figure 4.20 shows the number of types in the four examined corpora as a function of the size of growing corpus samples.

The number of different words within a corpus is usually interpreted as a measure of its lexical variance. The plot shows that the OpenSubtitles corpus had a much smaller vocabulary than the three other corpora which were clearly dominated by the blog posts in this respect.

PoS Distribution Table 4.19 lists percentage distributions for selected PoS tags on the level of tokens and types. We aggregated some of PoS categories for practical reasons. The figures show that the corpora were rather close in terms of tag distribution with a few remarkable differences. The higher amounts of pronouns and verbs in the subtitles is a direct consequence of shorter sentences. While the proportion of common names drops accordingly, this is not the case for the proper nouns, which validates the hypothesis that the subtitles actually replicate characteristics of spoken language. Besides, the lower proportion of common nouns and higher

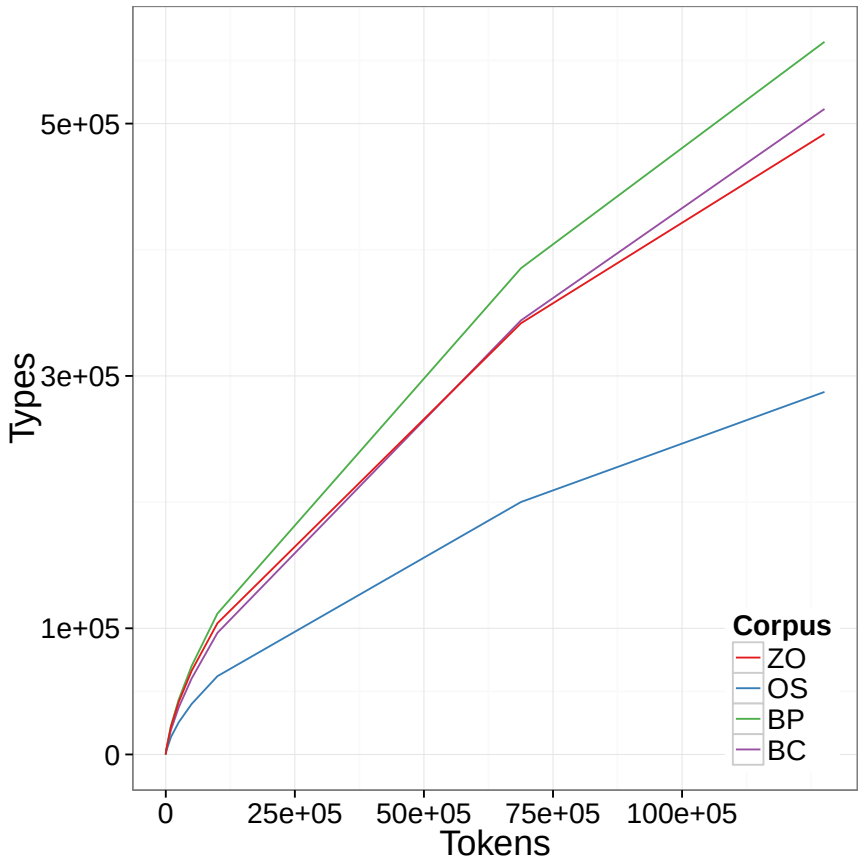


Figure 4.20: Number of types within random corpus samples (mean, 30 times iterated).

proportion of proper nouns in the blog comments indicates that it is relevant for studying vocabulary diversity.

Frequency Correlations For types shared by all evaluation corpora, Figure 4.21 shows correlations of their frequencies subdivided by frequency class. Frequency within the OpenSubtitles serves as the reference for frequency class since it is the largest corpus.

Correlations of subtitle frequencies with those from other corpora are clearly weaker than the other correlations while correlations of blog posts and comments are always higher. The general pattern was the same in all frequency classes but the differences between the single correlation values were smaller in the highest and lowest range.

Crps. / PoS	BP	BC	ZO	OS
Content words				
NN	16; 46	13; 42	18; 56	11; 42
NE	3; 22	2; 26	4; 18	3; 27
V*	12; 6	14; 8	13; 6	17; 9
AD*	14; 13	16; 14	13; 14	10; 11
Function words				
ART	8	6	10	5
AP*	8	7	8	4
P*	12	15	12	22
K*	5	5	4	3

Table 4.19: Percentage distribution of selected PoS (super)tags on token (content and function words) and type level (only content words). PoS tags are taken from the STTS. Aggregation of PoS categories is denoted by a wildcard asterisk. All percentages for function words on the type level are below one percent.

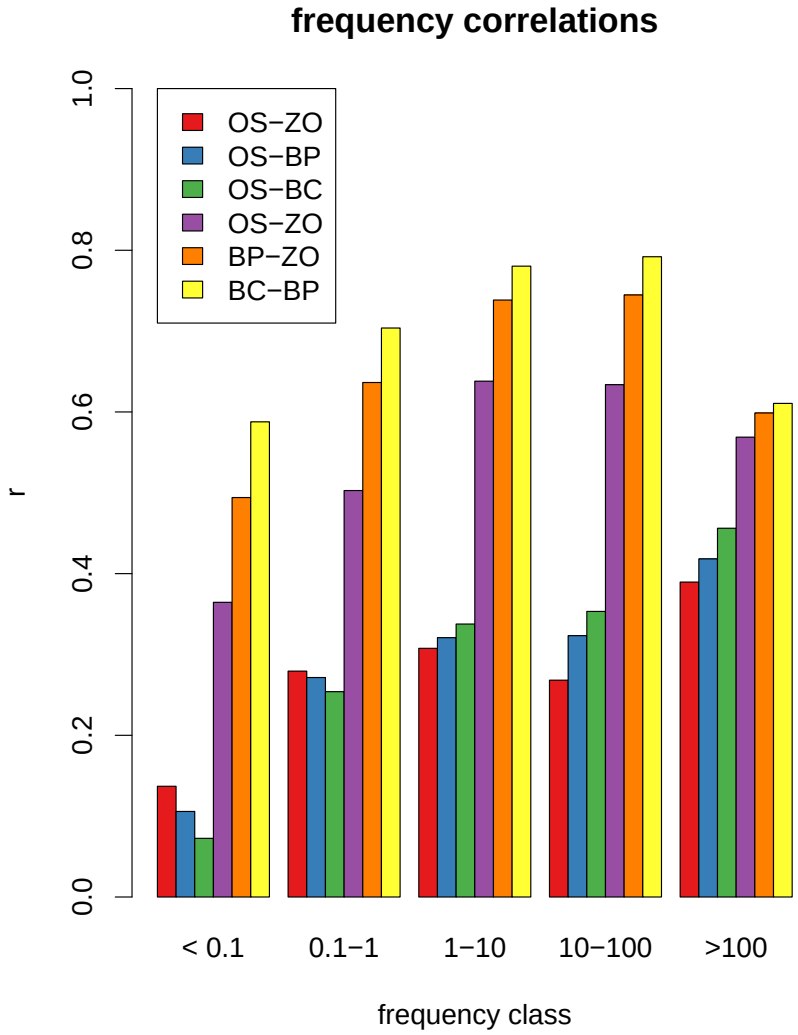


Figure 4.21: Correlations of type frequencies in different frequency classes.

Vocabulary Overlap Figure 4.22 shows overlaps in the vocabulary of the four corpora using a proportional Venn diagram (Venn, 1880). It was generated using the *Vennerable* (Swinton, 2009) **R** package which features proportional Venn diagrams for up to nine sets using the Chow-Ruskey algorithm (Chow & Ruskey, 2004). The diagram is arranged into four levels each corresponding to the number of corpora sharing a type. The yellow layer contains types which are unique to a certain corpus. Types shared by two corpora are mapped to light orange levels while dark orange levels contain types shared by three corpora. Types present in all four corpora constitute the central red zone. The coloring of the borders of the planes denotes the involved corpora. In order to abstract from the different size of the data sets involved and to allow for an intuitive comparison of the proportions within the diagram, we included only the 100,000 most frequent words from each evaluation corpus into the analysis.

Despite the heterogeneous nature of the corpora, there is a large overlap of roughly a third of the types between the four samples (red plane). Each sample contains a significant amount of exclusive tokens. The overlap between blog posts and comments is by far the largest on the second level while the one between blog posts and subtitles is the smallest. There is also a surprisingly large overlap between blog posts, comments and the ZEIT.

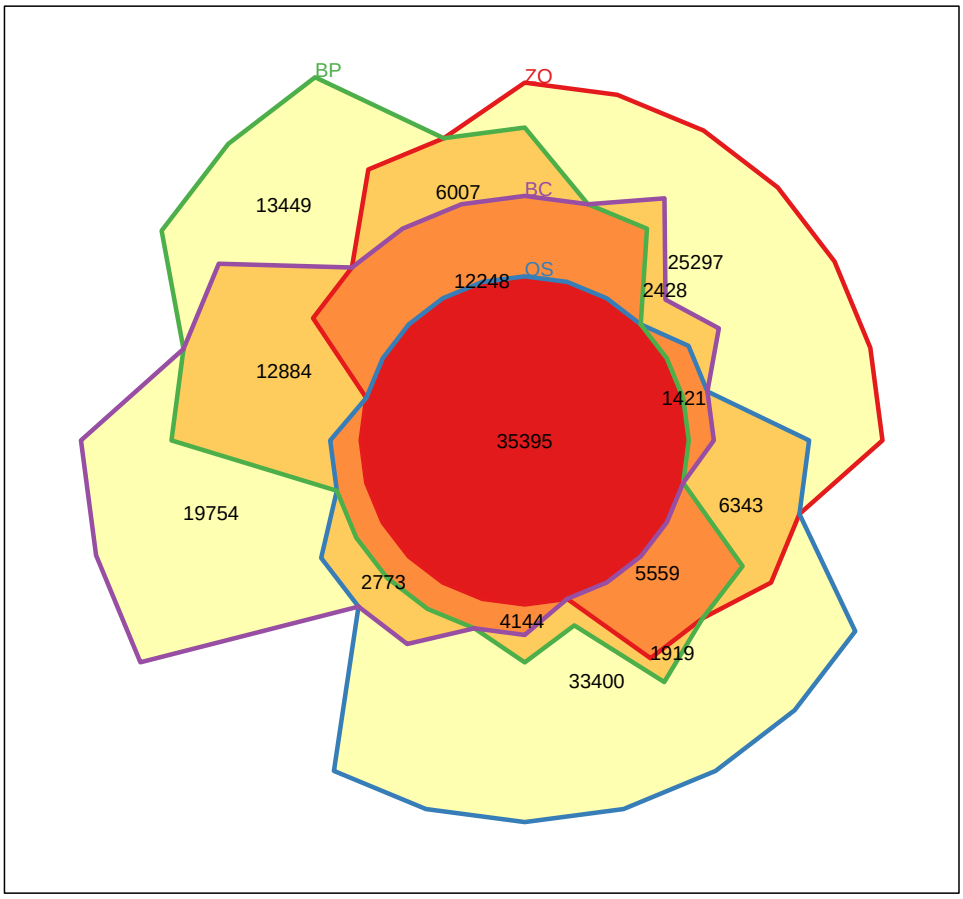


Figure 4.22: Venn diagram for the 100,000 most frequent words from each evaluation corpus.

Discussion The analyses above show large differences between the OpenSubtitles corpus on one and the ZEIT corpus on the other hand. These differences concern sentence length with much shorter sentences in the OS corpus; the amount of unknown words which includes non-standard word forms and (less frequent) named entities; frequency correlations which show large frequency deviations in the medium frequency range and PoS distributions with fewer nouns and more verbs for the subtitles. These results can be interpreted as resembling some of the differences between spoken and written language.

In almost all analyses, blog content was found to be closer to the ZEIT corpus than to the OpenSubtitles corpus. This might be expected for the posts but it was somewhat surprising concerning the comments which are to a great extent discourse-like communication. Nonetheless, the quantitative results were in accordance with qualitative results on that matter (Storrer, 2001; Dürscheid, 2003).

In exception to that pattern, the amount of tokens unknown to TAGH in the blog samples wa comparable to the value for the OpenSubtitles. This is caused by phenomena such as typos, standard-deviating orthography and *netslang* frequently observed in computer-mediated text and communication. In order to guarantee reliable linguistic annotation of blog posts and comments, emphasis will have to be put on improving existing and developing specific methods for automatic linguistic analysis.

Conclusion I have introduced strategies to try and classify blog corpora. Post content and comments seemed to be different in nature, so that there is a real interest in separate analyses, all the more since it is possible to perform text extraction and linguistic annotation efficiently enough to allow for a comparison with more traditional or established text types. In this regard, a corpus comparison gives insights on distributional properties of the processed web texts.

Despite the presence of atypical word forms, tokens and annotation UFOs, most probably caused by language patterns typically found on the Internet, token-based analysis of blog posts and comments seemed to bring these corpora closer to existing written language corpora.

More specifically, out-of-vocabulary tokens with respect to the morphological analysis were slightly more frequent in blog comments than in the other studied corpora. Concerning the lexical variance, blog posts dominate clearly, even if the higher proportion of proper nouns in the blog comments is a signal of promising richness for linguistic studies. Vocabulary overlap is best between blog posts and comments. However, a slight difference exists between them, the latter being potentially closer to subtitles, as the PoS tag distribution seemed to corroborate the hypothesis that subtitles are close to spoken language.

I believe that the visualization presented here can help to answer everyday questions regarding corpus adjustments as well as more general research questions such as the delimitation of web genres.

Future work includes updates of the resources as well as full downloads of further blogs. Longer crawls as well as attempts to access other blog platforms might be a productive way to build bigger and potentially more diverse transmissible corpora. Additionally, more detailed annotation steps could allow for a thorough interpretation.

Chapter 5

Conclusion

5.0.1 The framework of web corpus linguistics

The Web as chance for language resources Although it is not clear whether corpus linguistics can be fully established as a discipline, for the proponents of the framework – or discipline – it still has a promising future.

> "All in all, corpus linguistics can be argued to be a healthy, vibrant discipline within the general umbrella of language study. Its origins were non-computational but its explosion and expansion in the fields of descriptive and applied linguistics are due mainly to the information revolution of the late twentieth century, a revolution which continues, and from which CL [corpus linguistics] will undoubtedly continue to benefit." (McCarthy & O'Keeffe, 2010, p. 12)

In fact, web corpora are an example of how technological evolutions can change a research field.

I share the claim of Baroni et al. (2009) concerning the hope that other researchers will help "setting up what promises to be a pool of language resources among the largest ever made available to the research community". The Web offers the opportunity not to have to limit oneself to a ready-made corpus.[1]

Steps towards republishable content as well as open source software could foster further developments of research objects and methodology. In order to harness a changing technological environment without losing sight of linguistic goals and traditions, different skills are required, making cooperation within and between research teams in an interdisciplinary context mandatory:

> "It is our belief that the worldwide collaboration between corpus builders, as well as the cooperation between corpus builders and [...] engineers, will undoubtedly give rise to the further development of corpus linguistics." (Baker et al., 2004, p. 523)

However, I think it is necessary to take a stance on corpus linguistics so that corpus building and publication processes do not become issues merely related to corpus engineering. Corpus construction has to be established as a complex notion, involving theoretical and practical issues.

Additionally, changes within the discipline are calling to be interpreted and considered as part of a global evolution. All in all, (re)thinking corpora leads to a better understanding of intricate objectives and the constraints they are tied to.

New instruments, new observables, new science? Even from the point of view of science popularization, the change of scale of text collections is believed to introduce a change of perspective similar to the invention of the telescope or the microscope.[2] This is a common metaphor for referring to an evolution in linguistics linked to the general paradigm of technosciences and big data (see chapter 1).

[1] What Chevalier (1997) calls "un corpus préfabriqué."

[2] "From the perspective of a linguist, today's vast archives of digital text and speech, along with new analysis techniques and inexpensive computation, look like a wonderful new scientific instrument, a modern equivalent of the 17th century invention of the telescope and microscope."
Mark Liberman, *How big data is changing how we study languages*, in *The Guardian*, 7 May 2014.
http://www.theguardian.com/education/2014/may/07/what-big-data-tells-about-language

In that sense, even if it is not clear whether the change of scale is also a change of paradigm[3], it is certain that the existence of new instruments forces a change of the views on language as well as of the rules of these disciplines.

Moreover, these new instruments, with which it is possible to process bigger text and URL collections for instance, allow in turn for the creation of new observables (Valette, 2008, p. 11). The very first step may be to identify and register them rather than copy and transfer existing frameworks or theories onto them, in order to satisfy scientific constraints:

> "A science is not likely to be in a position to devise deep theories to explain its data, before it has an agreed scheme for identifying and registering those data"(Sampson, 2000, p. 1347)

Thus, even if there is for instance a tradition of genre theory going back to Aristotle's *Poetics* at least, the first step concerning web corpora seems to be a suitable characterization of texts, along with the proper labeling of phenomena. Additionally, the current text extraction and classification techniques were developed in the field of information retrieval and machine learning, i.e. not with linguistic objectives in mind. The whole situation makes it necessary to redefine, or at least specify the methods as well as the contours of corpus linguistics, which is attempted in chapter 1.

Even huge corpora are not neutral objects In the sense of linguistic tradition, and contrary to the popular belief that "more data is better data", it is useful to claim that corpora are not neutral objects simply waiting to be explored:

> "Data sets are not, and can never be, neutral and theory-free repositories of information waiting to give up their secrets." (Crawford et al., 2014, p. 1668)

In fact, corpus content is usually closely linked to its design and the series of decisions that led to the existence of a corpus.

To speak in the words of Bachelard (1927), it is true that the sheer mass of text lets general quantities and patterns emerge, as the "fine quantity disperses itself in a dust of numbers which cannot be registered anymore".[4] Bachelard stands for an aware and constructed execution of science and measurement technology, and a discovery process that does not search for precision itself or small details but assumes that approximation, without being too simple, is a key to knowledge acquisition.

However, results and byproducts of research on large corpora aren't neutral objects either. Design decisions constrain further developments. Even if it seems natural to derive patterns and regularities from web corpora, attention to their construction process is mandatory for drawing the right conclusions at the end of the toolchain. Moreover, the tools themselves cannot be expected to be systemically objective, so that both corpora and tools should be observed from a critical distance.

[3] "Peut-être constaterons-nous au final que l'intense activité tant en TAL qu'en linguistique aura surtout concerné une augmentation du volume et de la variété dans les données utilisées plus qu'un changement de paradigme." (Tanguy, 2013, p. 14)

[4] "Tandis que la quantité fine se disperse dans une poussière de nombres qu'on ne peut plus recenser, la quantité d'ensemble prend l'aspect du continu et se géométrise." (Bachelard, 1927, p. 154)

The linguistic tradition and the usual construction steps of digital corpora may be disrupted by the Web, and the history of debates on corpus balance. Representativity alone suffices in hinting at epistemological if not philosophical questions behind corpus creation.

> "Although in theory a corpus is a neutral resource that can be used in research from any number of standpoints, in practice the design of the corpus may strongly constrain the kind of research that is carried out. [...]
> Far from being neutral, then, issues of corpus design and building take us to the heart of theories of corpus linguistics. Questions of what goes into a corpus are largely answered by the specific research project the corpus is designed for, but are also connected to more philosophical issues around what, potentially, corpora can show us about language." (Hunston, 2008, p. 166)

I want to bring to (computational) linguists' attention that the first steps of web corpus construction seem to be underrated, although they influence many parameters at the core of the end corpus. Since the result is rarely known in advance, even with specialized web corpora, corpus evaluation and calibration is *a posteriori* rather than *a priori*, so that it is advisable to define "post hoc evaluation methods" (Baroni et al., 2009, p. 217).

Additionally, I agree with Baroni et al. (2009) that the ultimate test consists of how useful web corpora are to researchers in the field. In that sense, it is useful to create the right conditions to enable a hermeneutic circle in corpus creation and usage to function, with an identifiable user base and possibilities for dialogue.

5.0.2 Put the texts back into focus

Corpus quality and integrity Rather than size, the actual texts comprising the corpus should be put back into focus, particularly in terms of text quality.

The annotation of documents using ideally several metrics enable corpus designers as well as corpus users to decide which web documents to use during corpus building, corpus querying, or derivative production operations. In other words, corpus users should be put in a position to decide how important particular characteristics are for their purposes.

The corpus builder as a data gardener Internet data may be accessed directly, but the very notion of "corpora from the web" implies that important decisions have already been made along the way, along with a series of interventions during the processing steps. Additionally, the actual content of a corpus is often difficult to determine due to its size, and data cannot often be accessed directly, as it needs linguistic processing and a querying interface for instance. Thus, pitfalls and possible problems may not be easy to spot.

A great deal of work is needed to make research data be more accurate or conform to research objectives. This work often remains invisible, for instance unduly deleted texts or text parts, so that the work on corpus data can only be judged *ex negativo* by the quality of the end product.

In a nutshell, the corpus builder may be seen as a data gardener and sometimes even data janitor, as this part of corpus preparation is possibly the most prominent one in terms of time spent working on corpus material. The output of crawler and processing tools can leave more scraps and shreds of web content than actual texts or perfectly formed paragraphs. Thus,

separating the wheat from the chaff is a crucial operation to which research articles and reports should do justice.

Even statistical indicators can be affected by issues with content or metadata. The janitorial work is usually not given enough importance by the machine translation community for instance, where analysis by computational linguists can show that there are corpora that do leave room for improvement, e.g. Graën, Batinic, and Volk (2014) on the Europarl corpus.

Showing linguists that web corpora are no "disposable artifacts" Quality assessment of web corpora is necessary in order to convince more linguists to use them. A greater consideration for the whole processing chain could also avoid dismissing web corpora as being "disposable corpora" (Hunston, 2008, p. 158). If the Web is seen as a "mere text collection", it is most often used as a corpus by linguists, and for linguistic purposes.[5] That is why it is meaningful to establish web corpora as a reference and not only as a temporary resource.

Corpus exploitation in corpus linguistics means using a constructed view of language in a constructed discipline, as shown in chapter 1. It is also useful to bear in mind that not nature is observed, merely a garden. One may add that the evaluation of the "data scientists" and big data enthusiasts is not strictly of a linguistic nature, as it relies on complex tools such as the output of evaluation tools in statistical machine translation, where a mere numerical result is taken as input to justify or reject the validity of a whole approach or dataset.

Focusing on the texts could also help to prove that a web corpus is not an "artefact" which only represents "language on the Internet" (Hunston, 2008, p. 158). It is true that web corpora, no matter how large, cannot account for language use offline:

> "Big data continues to present blind spots and problems of representativeness, precisely because it cannot account for those who participate in the social world in ways that do not register as digital signals.(Crawford et al., 2014, p. 1667)

However, due to the progression of Internet use, language variants found online and offline are bound to coincide more and more. That said, the discussion on representativeness is not over, there is still some work to do on understanding as to what extent existing crawling strategies are biased, as well as determining how to establish a reliable sample, or at least a reasonable image of texts available on the Web.

Putting more web science into corpus construction As a plea for a web corpus creation aware of technicalities, I argue that a minimum of "web science" knowledge, as understood by Berners-Lee et al. (2006), in the corpus linguistics community could be very useful to fully comprehend all the issues at stake when dealing with corpora from the web. Altogether, page access delays, server-related biases, and unexpected web space topography are major issues impeding typical web corpus construction methods.

The use of search engines is not compulsory – Tanguy (2013) speaks of it as a "passage obligé". Thus, it does not have to be a bottleneck or an unstable source to take as is. It is truly convenient and does lead to content which is often relevant for building corpora.

[5] "Il faut donc considérer le Web comme, au mieux, une simple collection de textes dont la nature, la taille et la distribution sont mal connues (voire inconnaissables). Par contre, dans les usages concrets qui en sont faits dans les travaux de linguistique descriptive et de TAL, il semblerait qu'il soit utilisé en tant que corpus, au sens d'un ensemble de productions langagières exploitées à des fins linguistiques." (Tanguy, 2013, p. 9)

Light-weight, adaptable approaches Is the Golden Age of the Web as corpus behind us?[6] I do not think so, and I want to highlight that light-weight approaches are conceivable, where costs are lowered as much as possible, which might by a requirement for certain projects (Scannell, 2007).

My scouting approach using open-source software leads to a URL directory enriched with metadata useable for starting a web crawl. This is more than a drop-in replacement for existing tools since said metadata enables researchers to filter and select URLs that fit particular needs. They are classified according to their language, their length and a few other indicators such as host- and markup-based data.

A preliminary light scouting approach and a full-fledged focused crawler like those used by the Spiderling (Suchomel & Pomikálek, 2012) or the COW (Schäfer & Bildhauer, 2012) projects are complementary, since a preliminary phase can help to ensure a more diverse and less skewed sample distribution in a population of web documents, and/or to reach a given quantitative goal faster.

A light-weight scouting approach bears many advantages, such as easy updates, which have become crucial in a fast evolving Web:

> "Much data is dynamic, changing as new, better data is acquired or data treatment procedures are improved. There needs to be methods to ensure that linked databases can be readily updated rather than becoming 'stale'." (The Royal Society Science Policy Center, 2012, p. 65f)

All in all, the URL directory created here is part of a shift in common practices consisting of focusing on interesting seeds as well as part of the growing environment of linked databases.

Future work: better adaptation and specialization Work on text typology is needed both for general and specialized corpora. Further text characteristics and corpus descriptors may allow for a clearer classification of web texts and specification of research objectives as well as corpus types. Moreover, further indicators may provide easier access to corpus information for validation and assessment purposes, for example through the application of visualization techniques.

Concerning corpus specialization, more specific corpus types are conceivable, e.g. by distinguishing between several types of short messages, implying that inherent difficulties with respect to data gathering, processing, and republication have to be addressed.

Additionally, there is a growing need to adapt linguistic processing tools to the diversity of web texts and to enhance the robustness of the tools, since linguistic annotation can be crucial not only to the results but also to the corpus construction process:

> "Web crawlers need to be able to apply basic language technology such as tokenizers and language recognizers to the text in the web pages. For many languages, the simple forms of indexing used for English may not be sufficient, but some form of lemmatization and, e.g., decompounding may be necessary in order to build efficient information retrieval applications for the language." (Borin, 2009b, p. 10)

> "A second important line of research pertains to automated cleaning of the corpora, and to the adaptation of tools such as POS taggers and lemmatizers – that are often

[6] "Il est fort possible que l'âge d'or du Web comme corpus soit derrière nous." (Tanguy, 2013, p. 14)

based on resources derived from newspaper text and other traditional sources – to web data. Moreover, corpora should be enriched with further layers of linguistic annotation." (Baroni et al., 2009, p. 224)

The detection of content license is a good example of possible further annotation. Even if a manual verification is mandatory, distinguishing web pages under licenses which allow for republication, such as most Creative Commons licenses, may be a desideratum in order to progress towards free redistribution of web corpora. In general, work on licensing conditions is needed to clarify the imbroglio concerning copyrights and limitations of fair use.

5.0.3 Envoi: Back to the garden

Generally speaking, the two main types of web corpora, ad hoc and general-purpose, correspond to two different visions of collections and gardens, on one hand the world in a nutshell, and on the other a cabinet of curiosities.

Additionally, the idea of a garden helps to put the motives of corpus construction in perspective. To sum up, linguists require real, living samples of nature to study and on which to perform experiments. The rational desire and long-lived western tradition regarding the control of knowledge as well as of the architecture of the scientific process foster the ideal of corpora as miniature worlds, equipped with an overlook, and whose very structure is traversed by the notion of order.

Through the difficulty of establishing an order *a priori* and sometimes even *a posteriori*, web corpora challenge traditional notions and entanglements regarding text collections. They revive an different perspective on science and gardens: the impact of serendipity, surprise, and ultimately subjectivity. In that sense, evolving corpora and speaking trends on the Web resemble Gilles Clément's concept of "moving garden" (Clément, 1991), introducing freedom and unforeseen developments where there was traditionally order.

To conclude, let us not forget that "the luxury of gardens supposes that one loves nature".[7] Behind specific questions regarding texts and disciplines, there is usually a deep interest for language. Corpora are there to foster the acquaintances with objects of study[8], so that modern laboratories should not be soulless places. The joy of running experiments (*Experimentierfreude*) as well as the pleasure to dispose of a "cabinet of rarities" to show to others are legitimate drives of scientific discovery.

Further away from concrete expectations, the link between the place of experiments and the urge to strive for (inherent) beauty may have been forgotten, but approaches such as corpus visualization may be the ever so unconscious attempt to restore it.

It is my hope that this thesis has provided elements towards a better understanding of a research field as well as towards a middle ground between different traditions and motives. I share the belief of Friedrich Schiller, whose father Johann Kaspar Schiller was the entitled gardener at the court of the Duke of Württemberg, that a satisfying compromise can be found between "the stiffness of French taste" and "the lawless freedom of the so-called English one", which, in the original sense of mediocrity, may even be a very good solution:

[7] "Le luxe des jardins suppose toujours qu'on aime la nature." Germaine de Staël, *De l'Allemagne*, 1810, p. 15 http://gallica.bnf.fr/ark:/12148/btv1b86232882/f17

[8] In Tourtoulon's words: "la fréquentation de l'être à étudier" (Chevalier, 1997)

"Es wird sich alsdann wahrscheinlicherweise ein ganz guter Mittelweg zwischen der Steifigkeit des französischen Gartengeschmacks und der gesetzlosen Freyheit des sogenannten englischen finden; es wird sich zeigen, daß [...] es zwar abgeschmackt und widersinnig ist, in eine Gartenmauer die Welt einschließen zu wollen, aber sehr ausführbar und vernünftig, einen Garten, der allen Forderungen des guten Landwirths entspricht, sowohl für das Auge, als für das Herz und den Verstand zu einem charakteristischen Ganzen zu machen."[9]
Friedrich von Schiller, *Taschenkalender auf das Jahr 1795 für Natur- und Gartenfreunde*, *in* Allgemeine Literatur-Zeitung, Jena, 332, 1794, p. 99–104.

[9]My translation suggestion: "An acceptable middle ground can probably be found between the stiffness of French taste in garden design and the lawless freedom of the so-called English one; it can be shown that proposing to enclose the world within a garden wall is certainly vulgar and absurd, but that it is quite feasible and reasonable to bring a garden which answers all the demands of the good agriculturist to also form a characteristic entirety to the eye, the heart, and the mind."
https://de.wikisource.org/wiki/Taschenkalender_auf_das_Jahr_1795_f%C3%BCr_Natur-_und_Gartenfreunde

Chapter 6

Annexes

6.1 Synoptic tables

Register	Source	Quantification	Dates	Quality	Remarks
Newspaper	**Die Zeit**	1.000.000+ articles	1947-2014	very good	XML TEI
	Bild	100.000+ articles	1998-2010?	exploitable	raw text
	Der Spiegel	600.000+ articles	1949-2014	very good	XML TEI
Magazine	**Geo**	2.416 articles	2005-2012	very good	XML
Simplified magazine	**Geolino**	958 articles	?-2012	very good	XML
Simplified newspaper	**News4Kids**	300.000+ words	2004-2012	very good	XML
	Deutsch Perfekt	$\sim$ 20.000 words	?-2012	good	XML
Political speeches	**Chancellery**	1.831 speeches	1998-2012	very good	XML
	Presidency	1.442 speeches	1984-2012	very good	XML
Simplified dictionary	**HanisauLand**	$\sim$ 175.000 words	?-2012	very good	XML, CC BY-NC-ND
	Subtitles	10.000+ files	?	very good	raw text or XML TEI
Internet genre	**Blogs**	200.000+ files	?	very good	XML TEI, CC licenses

Table 6.1: Corpora created during this thesis

Name	Function	Availability
Bildcrawler	Crawling and processing for *Bild*	not released
Zeitcrawler	Crawling and processing for *Die ZEIT*	Google Code
Équipe-Crawler	Crawling and processing for *L'Équipe*	Google Code
GPS corpus tools	German political speeches corpus	Google Code
Diverse crawlers	for pages X and Y (Geo etc.)	not released
FLUX-toolchain	Light crawling scout	GitHub
Microblog-Explorer	Crawling and URL gathering on social networks	GitHub
Blog exploration	Crawling and processing of WordPress blogs	not released
LACLOS	Crawling and processing for OpenSubtitles	GitHub
URL-Compressor	Compression of large URL lists	GitHub
Toy crawler	Crawling experiments on a large scale	not released
Bloom filter & SQLite backend	Experiments on URL storage and retrieval	not released

Table 6.2: Tools developed in the course of the thesis

6.2 Building a basic crawler

This crawling approach has been used for the first versions of the news corpora, it started as a toy crawler and yielded results that were good enough to apply it in order to build larger scale corpora. Nonetheless, the intention here is not to introduce a new, particularly fast or reliable engine, it is rather to exemplify an exploring process among others. As the aim is to crawl specific pages of interest, it is based on particular knowledge of a website's structure.

The peculiarities of this example are taken from the French sports newspaper L'Équipe[1]. The scripts used were written in Perl in 2011. This kind of scripts work well for illustration purposes, but they are bound to stop being efficient as soon as the design of the website changes.

Settings First of all, one has to make a list of links so that there is something to start from. Here is the beginning of the script:

```
#!/usr/bin/perl   # assuming a UNIX-based system...
use strict;
use Encode;
use LWP::Simple;
use Digest::MD5 qw(md5_hex);
```

An explanation on the last line: we are going to use a hash function to shorten the links and make sure we fetch a single page just once.

```
my $url = "http://www.lequipe.fr/";
$page = get $url;
# because the pages are not in Unicode format
$page = encode("iso-8859-1", $page);
# taking the first eight characters of the md5 hash function of the url
push (@done_md5, substr(md5_hex($url), 0, 8));
push (@done, $url);
```

Now we have to find the links and analyze them to see if they are useful. Here, those with the word *breves* in it are of interest. A *brève* is a short description of something that happened, a few paragraphs long.

Gathering links

```
@links = ();
# taking the links one after another (not necessarily the fastest way)
@temp = split (" foreach $n (@temp) {
  if ($n =~ m/\/breves20/) {   # if the link contains the expression
     # absolute links
     if ($n =~ m/(http:\/\/www\.lequipe\.fr\/.+?)")/) {
     $n = $1;   # the first match
     }
  # relative links
```

[1] http://www.lequipe.fr

```
  else {
     $n =~ m/("\/?)(.+?)(")/;
     # the second group
     $n = "http://www.lequipe.fr/" . $2;
  }
  # just to check if the url looks good
  if (($n =~ m/breves20/) && ($n =~ m/\.html$/)) {
     # the URL list (or frontier)
     push (@links, $n);
  }
  }
}
```

If it is not the first time that the page gets fetched, one may want to check if one already went through that way using the first eight characters of the MD5 hash to spare memory.

Finally, one has to make sure there are no duplicates in the list so that it can be written to a file.

```
%seen = ();
@links = grep { ! $seen{ $_ }++ } @links;   # a fast and efficient way
```

Getting and cleaning text and metadata Then one has to go through the list one just made, a simple way is to get the pages one by one (since the bandwidth is the limit here it will not be necessarily slow).

First, one may want to charge the list of what one already did. One might also define a iteration limit for the loop that is going to start, so that one do not realize after a few hours that something in the script was not working properly.

Then the loop has to be started, the first page on the to-do list is retrieved. On can collect the remaining links on the fly as shown above.

Now one can find the author of the article, its title, its date and so on. I have an ugly but efficient way to do this: I cut off the parts that don't interest me so the informations are faster available using regular expressions. You can use splitting or regular expressions all the way, both work (at a certain cost).

```
### Cutting off
@temp = split ("<div id=\"corps\">", $page);
$page = $temp[1];
@temp = split ("<div id=\"bloc_bas_breve\">", $page);
$page = $temp[0];
### Getting the topic
$page =~ m/(<h2>)(.+?)(<\/h2>)/;
$info = "Info: " . $2;
push (@text, $info);
### Finding the title
$page =~ m/(<h1>)(.+?)(<\/h1>)/;
$title = "Title: " . $2;
push (@text, $title);
```

```
### Finding the excerpt if there is one
if ($page =~m/<strong>/) {
  $page =~ m/(<strong>)(.+?)(<\/strong>)/;
  $excerpt = "Excerpt: " . $2;
  push (@text, $excerpt);
  $page =~ s/<strong>.+?<\/strong>//;
}
else {
  push (@text, "Excerpt: ");
}
```

Remark: one could use XML fields as well. This example here is just a demonstration, it lacks functionality.

To get the text itself we could split the code into paragraphs, but it is not necessary here, as the HTML layout is basic. We just have to perform some cleaning of what we cut off, starting from the tags.

```
$page =~ s/<p>/\n/g;   #replacing paragraphs by newlines
$page =~ s/<.+?>//g;   #removing html tags
$page =~ s/SmartAd.+$//g;   #removing left ads
#... and so on
$page =~ s/^\s+//g;
#... and so on
```

Finally one writes the gathered text (here, `@text`) to a file and/or what one did and close the loop.

6.3 Concrete steps of URL-based content filtering

6.3.1 Content type filtering

```
# Main regexes : media filters
# avoid getting trapped
protocol = re.compile(r'^http')
extensions = re.compile(r'\.atom$|\.json$|\.css$|\.xml$|\.js$|\.jpg$|\.jpeg$|
\.png$|\.gif$|\.tiff$|\.pdf$|\.ogg$|\.mp3$|\.m4a$|\.aac$|\.avi$|\.mp4$|\.mov$|
\.webm$|\.flv$|\.ico$|\.pls$|\.zip$|\.tar$|\.gz$|\.iso$|\.swf$', re.IGNORECASE)
notsuited = re.compile(r'^http://add?s?\.|^http://banner\.|doubleclick|
tradedoubler\.com|livestream|live\.|videos?\.|feed$|rss$', re.IGNORECASE)
mediaquery = re.compile(r'\.jpg[&?]|\.jpeg[&?]|\.png[&?]|\.gif[&?]|
\.pdf[&?]|\.ogg[&?]|\.mp3[&?]|\.avi[&?]|\.mp4[&?]', re.IGNORECASE)
# avoid these websites
hostnames_filter = re.compile(r'last\.fm|soundcloud\.com|youtube\.com|
youtu\.be|vimeo\.com|instagr\.am|instagram\.com|imgur\.com|flickr\.com|
google\.|twitter\.com|twitpic\.com|gravatar\.com|akamai\.net|amazon\.com|
cloudfront\.com', re.IGNORECASE)
```

6.3.2 Filtering adult content and spam

The following Python-based regular expressions show how URLs which obviously lead to adult content and spam can be filtered using a rule-based approach.

```
## (basic) adult spam filter
# if ( ($testurl !~ m/[\.\/]sex|[\.\/-](adult|porno?|cash|xxx|fuck)/io) &&
($testurl !~ m/(sex|adult|porno?|cams|cash|xxx|fuck)[\.\/-]/io) &&
($testurl !~ m/gangbang|incest/io) && ($testurl !~ m/[\.\/-](ass|sex)[\.\/-]/io) ) {
...

# alternative 1
if re.search(r'[\./]sex|[\./-](adult|porno?|cash|xxx|fuck)', candidate) or
re.search(r'(sex|adult|porno?|cams|cash|xxx|fuck)[\./-]', candidate) or
re.search(r'gangbang|incest', candidate) or re.search(r'[\./-](ass|sex)[\./-]', candidate):
...

# alternative 2
if not re.search(r'[\./_-](porno?|xxx)', line.lower()) and
not re.search(r'(cams|cash|porno?|sex|xxx)[\./_-]', line.lower()) and
not re.search(r'gangbang|incest', line.lower()) and
not re.search(r'[\./_-](adult|ass|sex)[\./_-]', line.lower()):
...
```

6.4 Examples of blogs

The following examples taken on actual blogs under CC license show the variety of content layout as well as the difficulty to adapt to the diversity of the blogosphere, all the more since there are blogs for which it is useless to look for text content because it is inexistent or non-available.

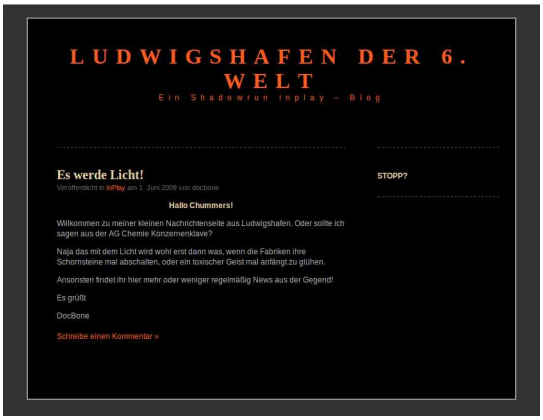


Figure 6.1: A new beginning... agchemludwigshafen.wordpress.com

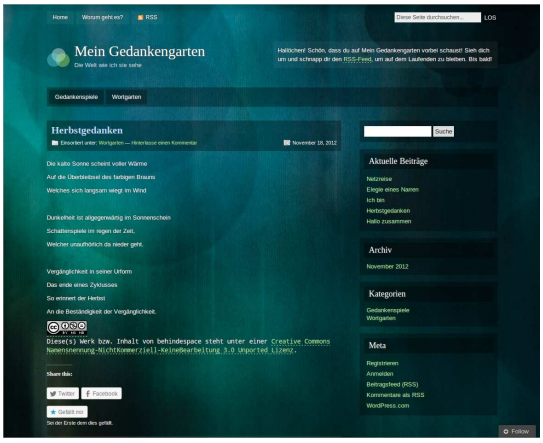


Figure 6.2: Poems behindespace .wordpress.com

Figure 6.3: Tweets I dmhdf.wordpress .com

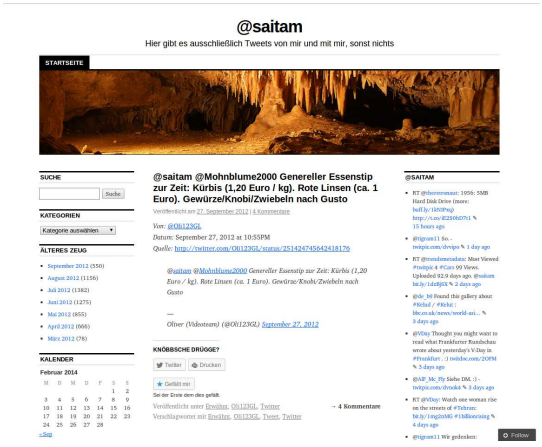


Figure 6.4: Tweets II saitam.wordpress .com

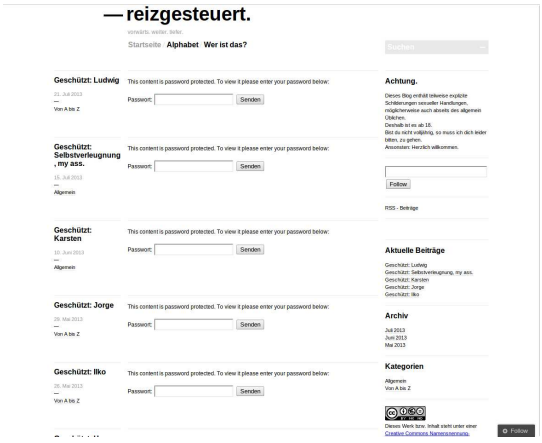


Figure 6.5: NSFW & NA reizgesteuert.wordpress.com

Figure 6.6: Mixed encodings blog.nihonnikonni.com

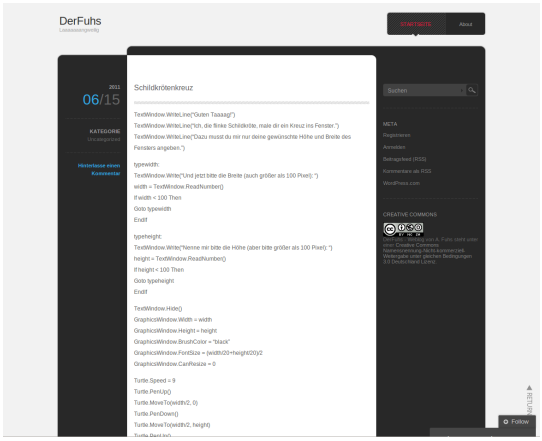


Figure 6.7: Code + natural language I derfuhs.wordpress.com

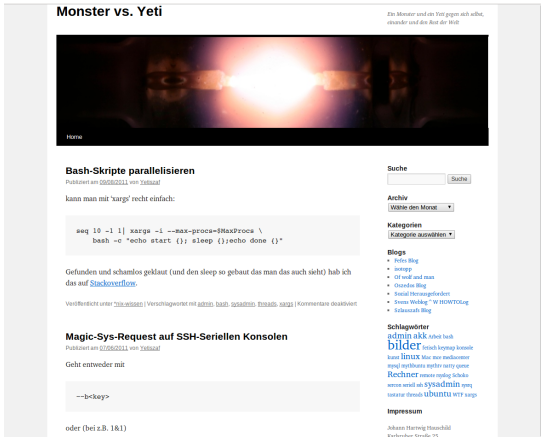


Figure 6.8: Code + natural language II blog.foxxnet.de

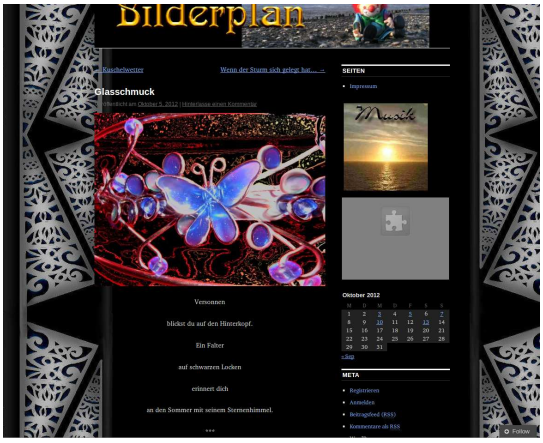


Figure 6.9: Visual effects I bilderplan .wordpress.com

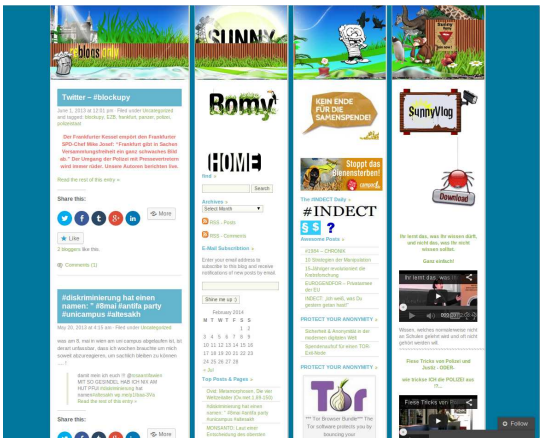


Figure 6.10: Visual effects II sunnyromy .wordpress.com

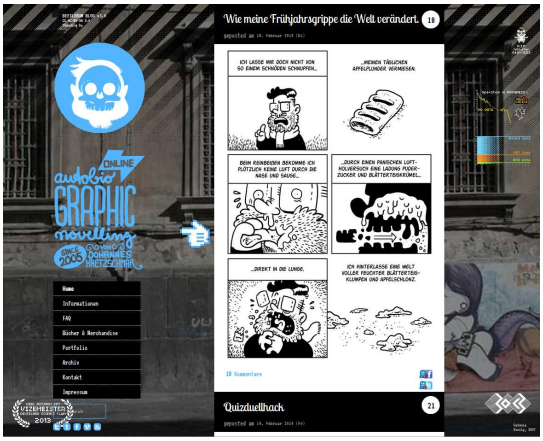


Figure 6.11: Comics blog.beetlebum.de

6.5 A surface parser for fast phrases and valency detection on large corpora

The following section has been published as ("A one-pass valency-oriented chunker for German", n.d.).

6.5.1 Introduction

6.5.1.1 Finite-state transducers applied to German

The idea to use finite-state automata to approximate grammar became popular in the early nineties, following the work of Pereira (1990) among others. As Karttunen (2001) reports, after a few decades of work on more powerful grammars due to the "persuasiveness of syntactic structures", computational linguists began working again with finite-state automata. The notion of chunk parsing (S. P. Abney, 1991) has been crucial for the evolution of finite-state parsers, as well as the notion of cascaded transducers.

The fact that these automata do not yield full parses but rather a series of indications obtained faster was considered to be particularly relevant, especially on the application side. Consequently the authors of the information extractor FASTUS stated that simple mechanisms can achieve a lot more than had previously been thought possible (Hobbs et al., 1997). As Neumann, Backofen, Baur, Becker, and Braun (1997) have shown, German is not an exception.

The growing interest in the research community towards the parsing using finite-state transducers of unrestricted texts written in German led to the publication of several mature parsers during the last decade. Kermes and Evert (2002) as well as Schiehlen (2003) use several levels of parsing to achieve a better precision, as they most notably enable to resolve ambiguities and to check the parsing structures for correctness. The finite-state approach proved adequate for German, as Hinrichs (2005) mentions: "It turns out that topological fields together with chunked phrases provide a solid basis for a robust analysis of German sentence structure".

The mature work on FST bears useful insights on the organization of German. For instance, FST parsers have problems with certain types of clauses, which is one reason why they were primarily dismissed by the advocates of generative grammar (Müller, 2007). Since Müller's doctoral thesis in 2007, little has been done to try to provide an overview of the state of the art, which may be explained by the efficiency of the parsers.

6.5.1.2 Practical interest of a valency-oriented tool

Given these abilities a less powerful approach could prove efficient when it comes to studying various syntactic phenomena, by using the strengths of the FST on one hand and exploiting the irregularities in the output from natural language processing tools, such as part-of-speech taggers, on the other in order to detect linguistic phenomena. In fact, non-finite state parsers have been found to provide helpful features but they are computationally demanding, and it can be interesting to see how far a finite-state approach is able to go when it comes to delivering fine-grained information.

Practical applications include readability assessment, isolation of difficult parts of a text, creation of selective benchmarks for parsers based on particular syntactical asperities as well as failure analysis. Hints can be used to assess text quality and/or quality of POS-tagger output, as the valency analysis reveals the existence of sentences without verbs or the lack of frequent

constituents such as head nouns for instance. This proves useful in non-standard text analysis, typically in learner or web corpora. Furthermore, it can also be used in these cases to assess the syntactical difficulty of a given phrase or sentence, which is considered an important criterion in readability assessment (Dell'Orletta et al., 2011).

This approach could also encompass what Biber (1992) calls "information packaging", stating that more detectable features linked to this notion could enable fuller models of discourse complexity. In a similar effort to combine different linguistic levels to get a more precise picture of text difficulty, Dell'Orletta et al. (2011) deal with "parse tree depth features", such as the depth of embedded complement chains and the number of verb dependents. They have taken over the research by Pitler and Nenkova (2008) who also used parser output features to detect syntactic complexity.

Thus, the use of the by-products of such tools to derive information about a text is common among researchers. However, the parsers employed in these studies are computationally complex, which makes analysis of large corpora dependent on time and resources. To our best knowledge it has not been tried so far to give an approximation for syntactic information, produced by simplified models designed on purpose.

My implementation of a chunk parsing method is part of annotation techniques designed to help qualify texts. More precisely, it is part of criteria which I documented in (Barbaresi, 2011a). These cues consist in a series of approximations of more sophisticated processes that are gathered in order to provide a "reasonable" image of text complexity. They are also a possible input for decision processes in web corpus construction (Schäfer et al., 2013).

6.5.2 Description

6.5.2.1 State of the art of this processing step

Several researchers have focused on this particular step, which is most of the time integrated in more complete processing tools. In the FASTUS approach (Hobbs et al., 1997), the basic phrases are such a step, where sentences are segmented into noun groups, verb groups, and particles. Another stage dedicated to complex phrases follows, where complex noun groups and complex verb groups are identified. The authors consider that the identification problems regarding noun phrases (such as the prepositional phrase attachment problem) cannot be solved reliably, whereas syntactic constructs such as noun groups can (i.e. the head noun of a noun phrase together with its determiners and other left modifiers).

My approach is also comparable to the segmentation part of the Sundance shallow parser (Riloff & Phillips, 2004) as well as to shallow parsing as shown by Voss (2005): the detection of indicators of phrase structure without necessarily constructing that full structure.

6.5.2.2 Characteristics of valency-oriented phrase chunking

The grouping into possibly relevant chunks enables valency detection for each verb based on topological fields, which is considered to be a productive approach for German grammar since the seminal work of Reis (1980).

The main difficulty criteria addressed by this approach are, on the intra-propositional side, the syntactic complexity of the groups (and possibly grammatically relevant phrases) and, on the propositional side, the complementation of the verbs as well as the topological nature of a phrase.

The transducer takes part-of-speech tags as input and prints assumptions about the composition of the phrases and about the position of the verb as output.

6.5.2.3 Characteristics of one-pass processing

Our approach aims at being robust. It takes advantage of the STTS tagset (Schiller, Teufel, Stöckert, & Thielen, 1995), and uses the tags as they are produced by the TreeTagger (Schmid, 1994). A more precise version including number, gender and case information provided by the RFTagger (Schmid & Laws, 2008) is possible and is currently under development. Nonetheless the newer tagger was significantly slower during our tests and thus it was not used for this study as it defeats the purpose of a one-pass operation in terms of computational efficiency.

The design is similar to parsers like YAC (Kermes & Evert, 2002), except that there is merely one step instead of several ones, as the program is designed to be an indicator among others. It deals with a linear approach, where the transducer takes one tag at a time without having to "look back", which accounts for computational efficiency.

The analysis relies on a pattern-based matching of POS-tags using regular expressions (which are themselves finite-state automata). The patterns take into account multiple possible scenarios of tag distribution. At each state, the transducer expects certain types of tags, which allow for a change of state. If the transducer starts, but does not find a given tag, it comes back to its initial state and ceases to output.

Forming a sort of ecosystem with the tagger, it is tightly dependent on it and requires to build on a stabilized one, whose decisions in common situations are (at least statistically) known.

Hand-crafted rules have already been considered as a noteworthy alternative to machine learning approaches (Müller, 2007). However, because of this fine-tuning, the chunker is limited to German.

6.5.2.4 Objectives

The purpose is neither to return a tree structure nor to deliver the best results in terms of accuracy (at least not primarily), but rather to yield various kinds of linguistic information useful to the language researcher.

The results are often comparable to text chunks, but the approach is closer to grammatical rules and to the definition of a phrase. The purpose is not to enfold every single particle, i.e. to achieve good recall, but to find word groups that are linguistically relevant with a good precision.

I share my objective with Voss (2005), which is to approximate a part of syntactic analysis that can be done automatically with few resources and glean as much syntactic information as possible without parsing.

6.5.3 Implementation

6.5.3.1 Detection of phrases

The detection of noun phrases and prepositional phrases takes place as shown in Figure 6.12. Starting from POS-tags following the STTS guidelines (Schiller et al., 1995), the transducer can go through several states and add tokens to the chunk according to certain transition rules before reaching its final step, i.e. a common or a proper noun (the tags NN or NE, respectively)

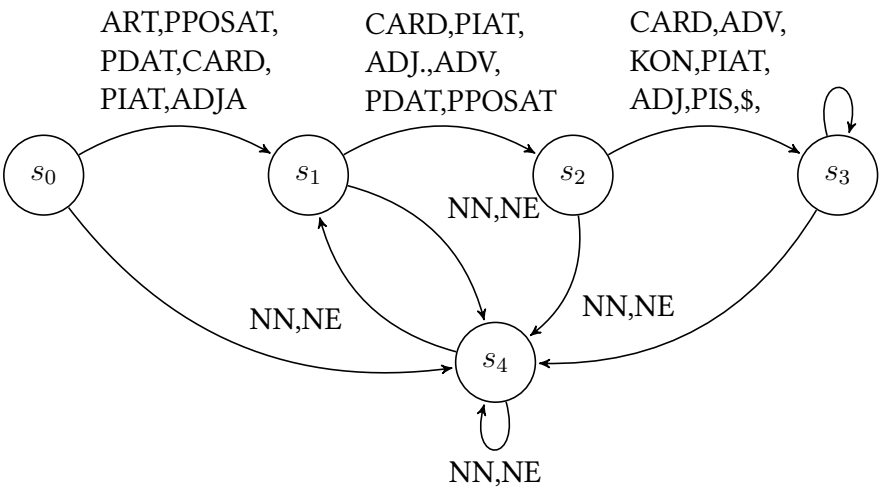


Figure 6.12: Simplified pattern used for detecting noun phrases on the basis of POS-tags using the STTS tagset. The additional APPRART and APPR tags are required to initiate the detection of prepositional phrases.

that is not followed by a word which could be possibly linked to the chunk, such as another noun or a tag which leads to the first state.

The detection of prepositional phrases is similar to mentioned scheme, with the main difference being the tags that allow a sequence to begin (APPRART and APPR). The head of the phrase is supposedly on the right of the group. The pattern is greedy: everything that fits under a predefined composition of a phrase counts. While this is a design decision that makes the implementation easier, it does not always perform well.

The chains of probable tags produced by the tagger as part of its operational design enable pattern analysis, which is based on known syntactical and grammatical rules, simple, well-known patterns, which as such are very likely to give satisfying results.

Thus, the constitution of the surface parser leaves little room for false incorporations, though abusive statements are not prohibited by design. Nonetheless, there is little chance of seeing incoherent output of the parser, since it takes benefit of the analysis by chains done by the tagger. The analysis of the tag probabilities given by the TreeTagger shows that there are two main cases: either it is quite confident about its output, or it fails at determining a reliable tag, which often affects several tags in a row. When the parser is confronted with such unusual tag chains, it ceases to output.

6.5.3.2 Actual valency

The purpose is to benefit from the detection mentioned above to give an estimation of the number of arguments that may be syntactically connected to a given verb. In order to do so, there are two operations needed, one on the extra-clausal and one on the intra-clausal level.

First, one has to find the boundaries of the clauses, since the sentence is not a relevant unit. In the case of German, this can often be done by locating the commas, as clauses are very frequently delimited by them, provided that they are not part of enumerations. Then, each head of a chunk found in a given clause increments the actual valency variable.

Due to the greediness of the phrase detection, the value is rather under- than overestimated,

Überfüllte	Einzimmerbehausungen	,	moderne	Apartments	oder	Kolonialvillen	im	französischen	Viertel	–	der
NP0	NP3		NP0	NP3	NP3-R	NP3-R	PP0-R	PP1-R	PP3-R		NP0
	1			1							
						1					

Fotokünstler	Hu	Yang	versucht	mit	seinen	Bildern	,	möglichst	viele	Facetten	seiner	Heimatstadt	einzufangen	.
NP3	NP3-R	NP3-R	VP	PP0	PP1	PP3			NP0	NP3	NP1-R	NP3-R	VP	
2						3				1				
1						**2**				**1**				

Figure 6.13: Example of the chunker output: Sentence text at the first level, the phrases being underlined, chunker output at the second level and valency counter at the third. The gold standard is at the fourth level in bold font. NP, PP and VP are phrase types, the numbers are states described in Figure 6.12. The letter R implies that an extension on the right has been detected.

which could prove interesting when it comes to comprehension level assessment. In fact, the estimated valency is most of the time between 2 or 5 complements per verb, which confers to a value of 5 or more a decisive character. In fact, this order of magnitude indicates that the sentence is bound to have a complex structure.

6.5.3.3 Proof of concept and adaptability

The transducer was first implemented to use it as a proof of concept, a standalone part of a text enrichment workflow. The code for this specific part has been made available online under an open source license[2]. As it is based on a series of conditional statements, IF-ELSIF-ELSE loops following the structure roughly pictured in Figure 1, it can be easily translated to another programming language. Other constraints can also easily be added. All statements can also be used in finite-state formalisms. The main dependency in terms of tools are the tagger and the tagset.

It is conceivable to use a “flat approach" of this issue by using one regular expression containing all the plausible scenarios and applying it directly to a whole text, a whole series of tags to match the candidates. Due to the computational complexity of long strings and multiple OR constraints, which is sometimes deteriorated by automata implementation issues of programming languages (Cox, 2007), this approach was not used for this study. The decomposition of the pattern and the use of a finite-state transducer benefits greatly to the processing speed as well as to the modularity of the analysis.

6.5.3.4 Example

The output of the finite-state transducer and the valency number guessed by the chunker is shown in Figure 2 under the text. “R" indicates a “greedy" right extension pattern was matched, the numbers indicate the state of the finite-state automaton as described in Figure 1), then the valency number guessed by the chunker. Finally, the numbers in bold font show the theoretically expected output.

The enumeration at the beginning is a problem, as it makes the proper identification of the base of the valency complementation a lot more difficult. The chunker fails at it, but still manages to count one complementation and not more, for instance because the commas are

[2]https://github.com/adbar/valency-oriented-chunker

used as a hint to detect an enumeration. This guess is false from a linguistic point of view, but by design a better precision cannot be achieved in those cases, as the end of the phrase comes unexpectedly late in the processing flow. This first part also illustrates the left-to-right parsing of the syntactic components, which could be overridden by a second pass. In this case, it is clear that one trades accuracy against this kind of robustness by adopting the one-pass approach.

The sequence starting with a dash shows a further problem, because in this case the counter should be reset. The noun phrases are identified properly, but the valency-complementation values are false.

The last part of the sentence is tagged properly and it shows the ability of the chunker to avoid issues link to the extensions on the right of the noun phrases (proper name and genitive adjuncts), to reset the counter at the beginning of a subordinate clause and to deal with discourse markers.

6.5.4 Evaluation

6.5.4.1 Large-scale analysis

Several grammatical particles are not taken into account, such as illocutionary and modal particles, adverbial portions of phrasal verbs and connectors.

Therefore, in order to evaluate the performance of the chunker, one can compute the ratio between the amount of tags that are concerned by the analysis and the amount of tokens for which there is no output. There are no evaluation metrics for the actual valency detection so far, since it is still an experimental feature which relies heavily on other processes.

The corpus used for evaluation consists of 2,416 recent online articles of the German version of the *Geo* magazine[3], comprising a total of 838,790 tokens. There are 469,655 non-verbal tokens for which there is an output and 234,120 verbal tokens (not only verbs but also modifiers like conjunctions or verbal particles) about which the transducer made a statement. Without the punctuation marks (representing 92,680 tokens according to the tags produced by the TreeTagger), that leaves about 6 % of the tokens that are possibly words without possible connections.

As already mentioned, the efficiency of the chunker regarding the particles it takes into account is interesting: 547,686 non-verbal tokens in total had a chance to be analyzed, which means that about 86 % of these tokens where considered to be part of a grammatically relevant chunk. If about 14 % of the tags were not incorporated, that means this information could be used to detect difficulties.

The cases for which there is no output are particularly interesting when it comes to text comprehension: if a grammatical structure is not recognized, then it may be a rare form or an error of the tagger. Both could be linked and both are relevant as a source of processing difficulty by humans or by machines. It can also mean that the structure is particularly long and/or complex, which is also relevant.

This information can also be used in order to isolate difficult parts of a text to compare the existing finite-state parsers, from which it is known that center embedding in noun phrases (Müller, 2007) or recursion issues are a source of problems.

[3] http://www.geo.de

6.5.4.2 Evaluation in detail

In order to give more precise insights on the performance of the chunker, its output has been evaluated on three different samples of 1,000 tokens in a row extracted from the corpus. The samples comprised a total of 180 sentences spread across eight different articles. The chunker found 831 valency complementations. 95 structures were falsely counted as valency elements, of which the noun phrase was correctly parsed but not numbered properly in 44 cases. 87 relevant structures were missed. Thus, the efficiency in terms of recall is slightly below 90 % on the test samples. The numeration accuracy is around 87 % and the F-measure for the values below is .890.

Output	Errors	Missed	Precision	Recall
831	95	87	.886	.894

6.5.4.3 Possible improvements

Close evaluation of the output has made it clear that there are two kinds of problems: those related to linguistics and those related to language processing issues.

On one hand, the impact in terms of valency of reflexive pronouns could be more adequately addressed. Trickier problems arise when loosely defined word categories come into focus, such as discourse markers, whose importance cannot be automatically verified using substitution tests. The task consisting of defining annotation guidelines based on acknowledged word categories in the field of linguistics is a challenge by itself.

On the other hand, a substantial part of the errors deal with tokenization and tagging artifacts such as falsely annotated URL components or punctuation issues. In fact, it is crucial in this one-pass approach to define a range of possible clause boundaries, from quotes to commas and to indirect speech markers, as it could improve precision.

6.5.5 Conclusion

A one-pass chunking and valency detection transducer has been presented. It is mainly linear and employs a bottom-up linguistic model implemented using finite-state automata. This allows by design for a fast processing speed and satisfies the constraints to work with large corpora.

Although design decisions can account for missing or false results in some cases, evaluation shows that this trade-off seems to be justifiable. There was an existing output for 86 % of the tokens in our corpus, and the valency counter's guesses are correct in 87 % of the cases. The first figure reveals that the chunker is quite permissive, whereas the latter shows that its accuracy is acceptable. Both metrics do not show what this tool could not integrate or analyze successfully, which is exactly where its possible application lies. This enables to focus on complex phrases or sentences as well as on irregularities in a corpus.

Future work includes three main topics of interest. First, an error analysis concerning on one hand the integration of certain grammatical particles and non-standard text-genres and tokenization artifacts on the other hand. Second, the integration of more precise morphosyntactic information which could enable a fine-grained analysis of the right extensions of the noun

phrase, for example genitive forms following nouns. The third topic of interest deals with metrics for actual valency detection, as the number of verbal dependents could be a highly relevant factor.

6.6 Completing web pages on the fly with JavaScript

Confronted with repetitive information, I looked for a way to make web pages lighter. JavaScript is helpful when it comes to save on file size. Provided that the DOM structure is available, there are elements that may be completed on load.

For example, there are span elements which include specific text. By catching them and testing them against a regular expression the script is able to add attributes (like a class) to the right ones. Without activating JavaScript one still sees the contents of the page, and with it the page appears as I intended. In fact, the attributes match properties defined in a separate CSS file.

I had to look for several JavaScript commands across many websites, that is why I decided to summarize what I found in a post.

First example: append text and a target to a link These lines of code match all the links that don't already have a `href` attribute, and append to them a modified destination as well as a target attribute.

```
function modLink(txt){
// Get all the links
  var list = document.getElementsByTagName("a");
  for (var i = 0; i < list.length; i++) {
// Check if they already have an href attribute
    if (!list[i].getAttribute('href')) {
// Get what's in the a tag
      var content = list[i].firstChild.data;
// Give the element a special href attribute
      list[i].href += "something/" + content + ".html" + "?var=" + txt;
// Give the element a target attribute
      list[i].target= "_blank";
    }
  }
}
```

The result: one has to call this function somewhere in the HTML document (I do it on load): `modLink(test;)`. Then, an a tag including the text 'AAA' will become a `href="something/AAA.html?var=tes` Very useful if one has to pass arguments.

Second case: append a class to a span element These lines of code modify existing span elements, those including parentheses and digits and the rest, by adding a class named c or i[digits] accordingly.

```
function modSpan() {
// Get all span elements
  var spanlist = document.getElementsByTagName("span");
  for (var n = 0; n < spanlist.length; n++) {
// Get their contents
    var inspan = spanlist[n].firstChild.data;
```

```
// If there is a left parenthesis
   if (/\(/.test(inspan)) {
// Find the digits in the contents and add them to a class named i
     var num = /[0-9]+/.exec(inspan);
     var add = "i" + num;
     spanlist[n].setAttribute('class', add);
// Hack to make it work with old versions of Internet Explorer
     spanlist[n].setAttribute('className', add);
   }
// If there is no left parenthesis in the span, add a class named c
   else {
     spanlist[n].setAttribute('class', 'c');
     spanlist[n].setAttribute('className', 'c');
   }
 }
}
```

The result: a span element containing text like 'AAA' will get a `class="c"` attribute, whereas another containing something like '(505)' will get `class="i505"`.

A last word regarding the security and the functionality of this code: it might be necessary to check the elements to change against fine-grained expressions (not like in this example) in order to ensure the result cannot be modified by mistake.

6.6.0.1 Display long texts with CSS, tutorial and example

I improved the CSS file that displays the (mostly long) texts of the German Political Speeches Corpus. The texts should be easier to read (though I did not study this particular kind of readability), here is an example.

I looked for ideas to design a clean and simple layout, but I did not find what I needed. Thus, I will outline in this section the main features of my new CSS file:

First of all, margins, font-size and eventually font-family are set for the whole page:

```
html {
margin-left: 10%;
margin-right: 10%;
font-family: sans-serif;
font-size: 10pt;
}
```

There are two main frames, one for the main content and one for the footer, denoted as `div` in the XHTML file.

```
div.framed {
padding-top: 1em;
padding-bottom: 1em;
padding-left: 7%;
padding-right: 7%;
border: 1px solid #736F6E;
```

```
margin-bottom: 10px;
}
```

I know there is a faster way to set the padding but I wanted to keep things clear and easy to maintain.

I chose to use a separation rule, `hr` in XHTML with custom (adaptable) spacing in the CSS:

```
hr {
margin-top: 2.5em;
margin-bottom: 2.5em;
}
```

This way title and text are much easier to distinguish.

Apart from the titles, which should be no mystery, another way to make the text look better is to justify and indent it, say for all paragraphs:

```
p {
text-align: justify;
text-indent: 1.5em;
}
```

6.7 Newspaper article in XML TEI format

The following code shows the final result of the extraction of article content for the newspaper corpus “Die ZEIT". The article taken as example[4] is made available as a XML file which complies with the guidelines of the Text Encoding Initiative. Due to copyright reasons most of the actual content of the article is stripped.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title type="main">Klimaschutz im Schnellspurt</title>
    <editor>
     <persName>
      <surname>Adrien</surname>
      <forename>Barbaresi</forename>
     </persName>
    </editor>
    <respStmt>
     <resp/>
     <orgName>Berlin-Brandenburgische Akademie der Wissenschaften</orgName>
    </respStmt>
   </titleStmt>
   <publicationStmt>
    <publisher>Berlin-Brandenburgische Akademie der Wissenschaften</publisher>
   </publicationStmt>
   <sourceDesc>
    <biblFull>
     <titleStmt>
      <title type="supertitle">Umweltminister Altmaier</title>
      <title type="main">Klimaschutz im Schnellspurt</title>
      <title type="subtitle">Nur sechs Stunden lang besuchte Umweltminister...</title>
      <author>
       <persName>Alexandra Endres</persName>
      </author>
      <editor>
       <persName>ldap.friese</persName>
      </editor>
     </titleStmt>
     <publicationStmt>
      <date type="publication">2013-11-20</date>
   <idno type="URL">http://www.zeit.de/wirtschaft/2013-11/altmaier-weltklimagipfel</idno>
      <idno type="UUID">8f28ef74-ce93-4253-a9d0-33b678a4c6d2</idno>
     </publicationStmt>
     <seriesStmt>
      <title>DIE ZEIT</title>
      <biblScope unit="year">2013</biblScope>
      <biblScope unit="volume">47</biblScope>
      <biblScope unit="ressort">Wirtschaft</biblScope>
     </seriesStmt>
    </biblFull>
   </sourceDesc>
```

[4] http://www.zeit.de/wirtschaft/2013-11/altmaier-weltklimagipfel

```
   </fileDesc>
   <profileDesc>
    <textClass>
     <classCode scheme="">online</classCode>
     <keywords>
      <term type="keyword"><measure quantity="8"/>Peter Altmaier</term>
      <term type="keyword"><measure quantity="6"/>Klimaschutz</term>
      <term type="keyword"><measure quantity="4"/>Connie Hedegaard</term>
      <term type="keyword"><measure quantity="4"/>Klimawandel</term>
      <term type="keyword"><measure quantity="3"/>Warschau</term>
      <term type="keyword"><measure quantity="4"/>Entwicklungsland</term>
      <term type="keyword"><measure quantity="4"/>Klimapolitik</term>
      <term type="keyword"><measure quantity="4"/>Minister</term>
      <term type="keyword"><measure quantity="4"/>Plenum</term>
      <term type="keyword"><measure quantity="4"/>Taifun</term>
      <term type="keyword"><measure quantity="3"/>Japan</term>
      <term type="keyword"><measure quantity="3"/>Philippinen</term>
      <term type="keyword"><measure quantity="3"/>Umweltminister</term>
      <term type="keyword"><measure quantity="1"/>Berlin</term>
      <term type="keyword"><measure quantity="1"/>Europa</term>
      <term type="keyword"><measure quantity="1"/>Paris</term>
     </keywords>
    </textClass>
   </profileDesc>
  </teiHeader>
  <text>
   <body>
<p>
...
</p>
<head rendition="#b">Emissionsziele</head>
<p>
...
</p>
<head rendition="#b">Klimafinanzen</head>
<p>
... Er soll ihnen eine Entschädigung für Klimaschäden garantieren.
<hi rendition="#i">Loss and damage</hi> heißt das in der Gipfelsprache. ...
</p>
<head rendition="#b">Streit über mögliche Entschädigungen</head>
<p>
...
</p>
<head rendition="#b">Deutsche Energiewende für internationalen Klimaschutz</head>
<p>
...
</p>
</body>
</text>
</TEI>
```

6.8 Résumé en français

Introduction

Cette thèse présente une réflexion théorique et pratique qui touche à plusieurs disciplines : la linguistique de corpus, la linguistique informatique et plus spécifiquement la construction de corpus tirés du web. Deux types de corpus web différents sont analysés, de même que les conditions nécessaires à leur compilation : les corpus spécialises d'une part, issus de sources ciblées et connues d'avance, et les corpus à usage général d'autre part, issues d'un parcours du web contenant une part d'aléatoire.

Chapitre 1

Contexte disciplinaire

Le premier chapitre s'ouvre par un description du contexte interdisciplinaire entre linguistique, linguistique de corpus et linguistique informatique. La linguistique informatique (ou computationelle) est une discipline relativement jeune, qui a connu des revirements théoriques et méthodologiques au fil des récentes évolutions de la technologie. La notion de linguistique de corpus, dont l'établissement en discipline fait débat, prend sa source – en théorie – dans un retour sur la notion de terrain et – en pratique – dans la possibilité d'explorer massivement des corpus de plus en plus grands via des outils informatiques.

Ce travail de thèse s'ancre dans ce cadre (inter-)disciplinaire, entre linguistique et linguistique informatique, avec d'une part une réflexion théorique sur des évolutions scientifiques et technologiques et d'autre part des incursions du côté des applications.

Ensuite, le concept de corpus est présenté en tenant compte de l'état de l'art. Dans une perspective synchronique, le besoin de disposer de preuves certes de nature linguistique mais embrassant différentes disciplines est illustré par plusieurs scénarios de recherche : les linguistes théoriciens, qui étayent leurs hypothèses avec des exemples tirés des corpus; les linguistes appliqués, expérimentaux ou de corpus, qui travaillent en rapport étroit avec les textes au format électronique, par exemple en effectuant des études quantitatives et en employant l'outil statistique afin de dégager des tendances; les linguistes informaticiens, qui interagissent directement avec les corpus en développant ou en entraînant des outils spécifiques, issus par exemple du monde de l'apprentissage artificiel ou des grammaires; les chercheurs en extraction d'information, discipline plus en lien avec l'informatique qu'avec la linguistique, où les corpus sont présents de manière indirecte en tant que données à traiter; et enfin des disciplines diverses dans le champ des humanités numériques, comme les sciences de l'information et de la communication, la sociologie ou l'histoire, où le web est perçu comme un objet culturel ou sociologique, ce qui change l'angle d'approche ainsi que les méthodes.

En somme, les différentes communautés décrites ci-dessus ne forment pas un ensemble homogène, leurs outils de même que leurs objectifs divergent, ce qui a une implication sur la manière dont les données de recherche sont rassemblées et étudiées.

Notion de corpus

Les définitions existantes des mots corpus et texte en linguistique sont discutées. Une définition large associe le corpus à une collection de documents, ici de textes, qui donnent un point de vue sur la langue.

Plus précisément, il existe une acception dite traditionnelle du corpus comme une collection organisée de textes soigneusement choisis. La notion traditionnelle de texte implique un certain nombre de conditions en terme de forme et de contenu (par exemple une nature discursive, une longueur de plusieurs pages, un style homogène). En conséquence, la construction de corpus selon cette tradition comporte notamment une première étape de spécification et de sélection des sources.

Les typologies classiques opposent les corpus généraux et les corpus spécialisés. Parmi les premiers, les corpus de référence ont pour but d'offrir l'image la plus raisonnable possible d'une langue en termes d'exhaustivité et d'équilibre de la composition. Au sein des typologies de corpus, leur exploitabilité ainsi que leur équipement sont également mentionnées, de même que la disponibilité des textes. Ainsi, d'autres critères viennent s'ajouter de manière transversale aux objectifs. Enfin, selon la tradition, il est important d'avoir une vision claire des différentes catégories de textes et de pouvoir délimiter des sous-corpus en fonction de leurs caractéristiques intrinsèques.

Dans ce contexte, les typologies existantes de corpus et de textes construisent de réelles différences et placent un certain nombre de contraintes entre les corpus d'intérêt général et les corpus spécialisés.

Étapes historiques de la construction de corpus

Dans une perspective historique, on peut distinguer plusieurs étapes clés de la construction de corpus, des corpus précédant l'ère digitale à la fin des années 1950 aux corpus web des années 2000 et 2010.

Le premiers travaux systématiques sur corpus à la fin des années 1950 incluent encore l'existence de fiches, ils ont une réelle importance dans la constitution de bonnes pratiques. Le corpus Brown, développé au cours des années 1960, est encore aujourd'hui un corpus qui fait office de référence concernant un certain nombre de critères constitutifs, par exemple concernant la combinaison d'avis d'experts et de hasard dans la sélection des textes ou le recours à des extraits de textes de longueur égale en raison des limitations des systèmes informatiques d'alors. Ces critères stricts ont pour objectif l'établissement d'un corpus en tant qu'objet scientifique, dans un contexte généralement hostile aux études empiriques et favorisant le développement de formalisations du langage.

La popularité croissante de paradigmes empiriques et des linguistiques de corpus émergeant dans les années 1980 est à replacer dans une tendance globale tirée par des évolutions technologiques vers une linguistique instrumentée et ce que l'on va appeler big data. À ce titre, la montée en puissance à la fin des années 1980 de textes qui sont dès leur naissance au format digital marque une césure. Les possibilités ouvertes par les avancées technologiques dans le domaine de l'informatique s'inscrivent dans un contexte global d'une linguistique toujours plus instrumentée, et au-delà d'approches inductives dites tirées par les corpus, où le terrain n'est plus seulement un espace où vérifier des hypothèses mais un masse critique à même d'en suggérer. Le positionnement de la linguistique de corpus manquant de clarté s'ajoute à l'existence simultanée de plusieurs méthodes de recherche, divisant la communauté scientifique. Le corpus devient un mélange d'objectifs a priori et de contraintes liées au(x) traitement(s).

La tradition établie de facto évolue encore de manière significative avec l'arrivée des données tirées du web. L'évolution des politiques scientifiques vers ce qu'il convient d'appeler les technosciences ainsi que la part croissante de l'interdisciplinarité joue constamment un rôle à l'arrière-plan. De plus, l'arrivée des mégadonnées questionne les pratiques existantes, au sens

où certains estiment désormais que les données se suffisent á elles-mêmes, et qu'il faut inclure toujours plus de données. Les continuités et changements entre la tradition en linguistique, c'est-à-dire les corpus de référence, et les corpus tirés du web sont visibles notamment sous l'angle de la représentativité ainsi que de la praticabilité du résultat. Les raisons techniques et idéologiques d'un départ du paradigme envisageant le *web comme corpus* en découlent, de même que des arguments en faveur du *web pour les corpus*.

Changements liés aux textes tirés d'Internet

Le premier chapitre se conclut sur la montée de la linguistique opérant sur des corpus extraits d'internet, et les changements que cette approche induit dans leur construction. En effet, plusieurs critères d'établissement des corpus classiques posent problème dans le cas des corpus web, comme par exemple la notion de texte (à la fois du point de vue des textes courts et de la difficulté à séparer texte et paratexte), la notion de représentativité (la représentativité d'une page web étant parfois difficile à quantifier ou à comparer), les droits concernant les textes (par la percée des licences libres mais aussi des droits d'auteur plus complexes et internationaux).

Les questions abordées au cours de la thèse sont soulignées : la qualité et l'exploitabilité des données, une construction qui puisse être accessible, indépendante et pratique, ainsi que le croisement des approches entre plusieurs traditions scientifiques et plusieurs disciplines, en France comme en Allemagne et ailleurs.

En somme, on peut dire que la linguistique de corpus web est en plein essor, malgré un certain scepticisme que l'on peut relativiser car les problèmes posés par les textes tirés du web tiennent en partie à une définition trop floue, changeante ou sujette à controverse de la linguistique de corpus ou même des notions fondamentales de texte et de corpus. Parmi les principaux changements introduits, on peut citer les difficultés dans la normalisation et la classification des textes.

Chapitre 2

Le second chapitre rassemble des considérations méthodologiques sur l'examen automatique de textes en informatique, linguistique informatique et traitement automatique des langues.

Estimation de la qualité des textes

Premièrement, l'état de l'art concernant l'estimation de la qualité de textes est décrit en guise d'exemple de recherche interdisciplinaire sur des textes venant d'internet. Quelques principaux problèmes tels que le spam et leurs (imparfaites) solutions montrent l'importance de trouver des indices pour des phénomènes qui frappent l'œil humain sans être détectés par les machines.

Plus précisément, le cas des textes générés automatiquement, des textes traduits automatiquement, du spam rédigé à la main et de la présence de plusieurs langues sur la même page web sont relativement fréquents et font partie des problèmes soulevés par les corpus web. C'est pourquoi les acquis de la recherche en informatique sur ces thèmes peuvent être transférés pour améliorer la qualité des corpus.

De plus, les machines peuvent elles-mêmes faire l'objet de diverses manipulations, qui ont par exemple pour but d'améliorer le classement d'un site web dans les moteurs de recherche. En ce sens, l'existence de contenu nativement digital et traitable par des machines a des limites

qu'il faut considérer : sans éviter les pièges destinés aux machines, il est possible de trouver des textes qui n'ont rien à voir avec l'usage de la langue par les hommes.

Enfin, la reproductibilité du texte à l'heure d'internet a aussi un effet certain sur la composition des corpus, en l'occurrence par la présence de nombreux doublons. Des critères d'adoption des textes issus d'une approche interdisciplinaire sont nécessaires pour améliorer la qualité des textes, ils peuvent être relativement simples, comme un critère de longueur des pages et des textes pour la détection des doublons, ou plus complexes, comme des modèles de langue pour la détection du texte généré par des machines.

Études de lisibilité

Ensuite, les méthodes utilisées par les études de lisibilité ainsi que par la classification automatique de textes peuvent servir de parangon de l'identification de caractéristiques textuelles pertinentes, et ce de la recherche théorique aux applications industrielles.

Ainsi, les étapes dégagées par la classification de textes (établissement d'un étalon, définition d'un ensemble de prédicteurs et calibrage d'un sous-ensemble de critères utilisés pour l'apprentissage artificiel) peuvent servir à déterminer la lisibilité des textes en particulier et l'adéquation de textes à un standard en général.

De même, les différentes approches utilisées par les études de lisibilité illustrent différents rapports aux textes et différentes approches dans leur examen. Ainsi, la linguistique, la psycholinguistique, les sciences de l'éducation, les approches industrielles et la linguistique informatique ont des dénominateurs communs qui donnent une bonne indication de l'état de la recherche dans ce domaine.

Par exemple, la surprenante efficacité des critères de surface semble commune à plusieurs disciplines. De même, le rapport entre microstructure et macrostructure, complexité(s) locale(s) et complexité globale, indiquent que différentes échelles sont nécessaires pour appréhender ce phénomène.

On peut retenir de l'état de la recherche de l'estimation de la qualité et de la lisibilité qu'il n'y a pas de large consensus en ce domaine mais plutôt des avancées ponctuelles. Cela dit, la qualification de textes tirés du web semble être un domaine d'application potentiel des études de lisibilité, sujet abordé au cours du chapitre 4.

Visualisation de textes

Enfin, la visualisation de textes démontre le potentiel intérêt de l'analyse de corpus pour les humanités numériques. Le principe de visualisation globale, qui procure un apercu à l'échelle du document, s'oppose à la visualisation à une échelle plus restreinte, qui permet de donner plus de détails au prix peut-être de la simplicité d'interprétation.

En général, en linguistique de corpus, les techniques de visualisation proposent d'explorer un ou des corpus, de manière interactive ou non, afin d'apporter un autre regard sur des masses de textes trop vastes pour l'œil humain, selon le concept de la lecture distante.

En conclusion, quelques bons principes visant guider la recherche sont annoncés concernant l'importance des règles d'annotation, des facteurs de surface, d'une approche globale des textes et concernant l'impact de la visualisation. Étant donné que la majeure partie des techniques exposées dans ce chapitre appartiennent à la sphère des sciences appliquées, si bien qu'un équilibre entre analyse quantitative et linguistique de corpus doit être trouvé. Corpus et indicateurs restent des constructions dont il garder à l'esprit le caractère artificiel et

l'abstraction qu'ils représentent par rapport à la réalité qu'ils sont censés décrire ou représenter (la langue pour le corpus, les textes pour les indicateurs).

Chapitre 3

Le troisième chapitre résume l'apport de la thèse en ce qui concerne la recherche sur les corpus tirés d'internet. Il s'ouvre sur des notions de *science du web* et un aperçu des contraintes liées au corpus non dynamiques. Le web commence à être étudié dans le cadre de ce que l'on appelle la science du web, cependant son impact n'est pas encore vraiment théorisé ou étudié dans le cadre des sciences humaines.

URLs et web crawling

Les URLs sont un composant essentiel qui confèrent au web robustesse et permanence. Pour décrire sa forme, les chercheurs en informatique recourent fréquemment à l'image d'un graphe avec des nœuds et des arêtes. La distribution des sites web au sein de cet ensemble suit probablement une loi de puissance, et la distribution des liens entre les sites n'obéit pas au hasard. En conséquence, on peut considérer que le web est une structure polynucléaire ou les noyaux sont plutôt denses et bien reliés, avec une vaste périphérie de sites disséminés deci delà.

La notion de web crawler est centrale pour comprendre comment les corpus web sont construits. Par web crawling on entend un parcours du web par des machines, qui téléchargent et le plus souvent archivent des pages web en série. En raison de la prévalence des modèles informatiques de graphes, un crawl est considéré comme l'opération consistant à traverser ce graphe qu'est le web.

Lors de la conception de corpus web dits hors-ligne, il s'agit donc tout d'abord de trouver des sources de contenu, de sélectionner les documents pertinents, de fournir des métadonnées, d'ôter le bruit, le spam et autres parties inintéressantes pour un chercher en linguistique, et enfin de republier l'ensemble, ou tout du moins de le rendre accessible.

Trouver des sources pertinentes

La question de la collection des données est examinée avec une attention particulière, tout spécialement le cas des URLs sources. Deux principales approches sont distinguées, d'une part le cas de documents déjà listés ou connus, et d'autre part la découverte de documents web, qui si elle est complexe devrait être incluse dans le champ de la linguistique, tant elle a d'implications pour les études conduites en aval.

Les URLs sources sont des URLs utilisées pour démarrer une phase de web crawling. Il est crucial pour le déroulement du crawl que ces URLs soient pertinentes en termes de contenu ciblé, par exemple en termes de langue(s) utilisée(s). Dans le cas de documents déjà connus, une liste d'URLs sources équivaut au nombre de pages que l'on veut archiver, tandis que dans le cas de corpus généraux, elle n'est que la première étape, la partie émergente du crawl.

La méthode BootCaT fait figure de référence dans la création de corpus web généraux. Elle implique d'utiliser des moteurs de recherche et des séries de mots choisis au hasard afin de découvrir des URLs. Si cette approche s'avère productive, elle est toutefois sujette à caution. En effet, sa simplicité masque un certain nombre de questions quant à son opacité, essentiellement due à la dépendance vis-à-vis des moteurs de recherche. Ces derniers sont opérés par des sociétés transnationales répondant à des logiques commerciales et non scientifiques, ce qui

signifie d'une part que la gratuité des requêtes et la pérennité des résultats ne sont pas garanties et que d'autre part les réponses fournies peuvent varier considérablement dans le temps ou selon le profil attribué à l'utilisateur.

D'autres facteurs problématiques sont à relier aux changements du web lui-même, par exemple la popularité croissante de pages dites dynamiques et du paradigme du web 2.0, qui signifie pour l'archiviste qu'une page peut changer à tout moment, que ses sources sont multiples et donc difficilement démêlables ou attribuables automatiquement, et enfin qu'une URL n'est qu'une voie d'accès à des contenus mobiles qui déjouent la rigidité de la notion d'URL.

Crawling déterministe et crawling sauvage

Le rassemblement de documents déterminés d'avance n'est pas concerné par ses évolutions, étant donné que même dans le cas de multiples sources le contenu passe par une phase de vérification ou de curation par les concepteurs du corpus ou du portail utilisé.

En revanche, dans le cas du web crawling en pleine nature, le résultat ne peut être connu d'avance et un grand nombre de facteurs proprement techniques influent sur le déroulement des opérations. On pourrait ainsi penser qu'il s'agit d'un processus trop complexe ou de nature trop technique pour être conduit par des linguistes. Mais précisément, la construction de corpus web devrait être maîtrisée du début à la fin, et ce par ses instigateurs, même si cela implique de développer des connaissances en science du web et de suivre en détail des applications logicielles. En effet, l'impact du web crawling sur la composition du corpus est tel que cette étape devrait être mieux comprise et avoir plus d'importance.

On peut distinguer quatre étapes au sein de ce processus. Premièrement, les URLs sources sont découvertes et rassemblées. Deuxièmement, le contenu cible est déterminé (type, langue, caractéristiques techniques, etc.) Troisièmement, il faut se positionner en termes de politesse du crawl (décider d'outrepasser certaines bonnes pratiques ou non). Enfin, le crawl peut commencer, ce qui implique éventuellement de suivre son déroulement.

Il existe différentes stratégies de parcours du web, les deux principales étant le crawl en largeur, incluant le plus de sites possibles au fur et à mesure qu'ils sont découverts, et le crawl restreint, qui cible de manière précise une langue, des composantes d'URL, ou un sujet précis. En tout cas, il est très souvent impossible d'être absolument exhaustif. Les principales limites sont la taille du web, les documents inaccessibles, les changements d'URLs et l'évolution rapide des pages qu'elles référencent.

Afin d'être plus précis dans la conception du corpus, il est possible de prendre pour cible une part très réduite du web, une communauté de locuteurs relativement restreinte, en définissant un certain nombre de contraintes, comme la présence majoritaire d'une langue rare et/ou un type de communication précis comme les microblogs. Cependant, malgré les caractéristiques de surface apparemment homogènes et les attentes en ce sens, la variabilité des résultats demeure. Comme ces domaines de recherches sont encore jeunes, il n'existe pas de consensus en ce qui concerne la découverte de pages pertinentes. La technique la plus raisonnable consiste à maximiser les chances de trouver ce que l'on cherche même avec peu de moyens. Des métadonnées riches sont également un moyen de rendre le résultat plus exploitable.

Pré- et post-traitements

La notion de pré-traitement des corpus web est introduite, ses étapes majeures telles que le filtrage et le nettoyage sont brossées. L'impact des pré-traitements sur le résultat est évalué à

la lumière de réels exemples, afin de montrer pourquoi ils ne devraient pas être sous-estimés.

En effet, la qualité du corpus peut varier fortement non seulement en fonction du déroulement d'un crawl, mais aussi en fonction des traitements appliqués en vue de sa création. En ce sens, il est sans doute préférable de parler de pré-traitements plutôt que de post-traitements comme le fait une partie de la littérature. Une part significative des documents est rejetée, le choix des logiciels influence le résultat, et enfin le préfixe pré- montre clairement qu'il ne s'agit pas d'une étape facultative mais bien d'une étape clef.

Les pré-traitements impliquent tout d'abord de cerner les parties pertinentes des documents et de rejeter le reste. À cette fin, le balisage HTML disparaît, et le paratexte doit être séparé du contenu principal des pages, ce qui les rend uniques. De plus, dans certains cas, des documents entiers sont rejetés parce qu'ils ne correspondent pas aux objectifs de la création de corpus.

Une partie de ces traitements peut se faire sur la base de critères relativement simples, comme la taille du document et certaines caractéristiques de surface. Concernant la pertinence des textes, plusieurs approches existent, certaines à base de règles, d'autres utilisant l'apprentissage artificiel.

Enfin, l'inclusion proprement dite des textes dans le corpus suppose un dernier examen pour déterminer au mieux la structure de l'ensemble. Cet examen des caractéristiques globales inclut par exemple l'estimation de la part de doublons et leur éventuel rejet.

Dans l'ensemble, l'impact des pré-traitement est significatif, puisqu'il n'est pas rare que plus de 90% des textes soient rejetés au cours de la procédure afin d'améliorer le résultat final. Cela dit, les corpus web contiennent par nature toujours une certaine proportion de déchet qui a souvent un impact qualitatif plus que quantitatif.

Comme le parcours du web par des machines est simple en apparence seulement, la question de la simplicité et de la reproductibilité de la construction de corpus doit être mise en avant. Si un certain opportunisme est nécessaire pour s'intéresser aux corpus web, une approche scientifique est nécessaire afin d'établir ces corpus comme des objets construits scientifiquement, d'après un état de l'art et des techniques, des comparatifs des solutions logicielles, et une traçabilité des intentions des créateurs. Une partie des questions soulevées par cette démarche est examinée dans le chapitre 4.

Chapitre 4

La quatrième partie décrit l'apport de la thèse du point de vue de la construction de corpus proprement dite, et ce sous deux principaux aspects : la question des sources (concept de pré-qualification) et le problèmes des documents invalides ou indésirables (concept de qualification).

Hypothèses

Tout d'abord, il est prouvé qu'il est possible et même désirable d'utiliser des sources autres que simplement les moteurs de recherche pour trouver des documents. En ce sens, une approche utilisant un éclaireur léger pour présélectionner des pages web et préparer le parcours du web est désirable. Pour ce faire, il s'agit de trouver des caractéristiques textuelles qui correspondent aux particularités des textes tirés du web. Dans un second temps, les URLs (et le cas échéant les documents web) doivent être filtrés et annotés selon les besoins des chercheurs en aval. Enfin, les documents téléchargés doivent être analysés quant à leur possible inclusion dans un corpus.

À cette fin, l'état de la recherche présenté dans le chapitre 2 peut être utilisé pour développer des filtres et des métriques d'évaluation de la classification des URLs (ou pré-qualification) et des documents web (ou classification proprement dite). Enfin, en utilisant des techniques d'analyse et de visualisation, il est possible d'obtenir une vue d'ensemble d'un ou de plusieurs corpus.

Pré-qualification

La suite d'outils FLUX est une approche en surface de la pré-qualification qui a pour but d'être plus légère et de requérir moins de ressources qu'un crawler complet. Elle prend en entrée une liste d'URLs et propose en sortie une liste d'URLs viable, c'est-à-dire plus prometteuse que des URLs prises au hasard en ce qu'elles ont été testées en regard d'une série de caractéristiques pertinentes pour la création de corpus.

La première étape de la suite d'outils consiste en la collecte d'URLs, via des requêtes, des sites annuaires ou l'analyse de sites comme les réseaux sociaux. Cette étape inclut également le filtrage des spams ou des URLs dont on peut dire a priori qu'elles ne sont pas pertinentes. Ensuite, la liste d'URL subit une première analyse avec notamment des tests de redirection et un échantillonnage par nom de domaine. Enfin, les documents web pointés par les URLs actualisées et regroupées par ce biais sont téléchargés et analysés afin d'obtenir des métadonnées et de permettre leur classification.

Exemple 1 : le cas des langues rares Des tests menés sur plusieurs langues plus ou moins rares (indonésien, malais, danois et suédois) ainsi que sur plusieurs sources (l'annuaire DMOZ, Wikipédia et des portails thématiques) montrent qu'il est réellement utile de contrôler les URLs afin de retenir les plus productives. De plus, des sources peuvent tout à fait être trouvées en-dehors des moteurs de recherche. Les métadonnées indiquent qu'une conception de corpus à partir de ces sources peut-être envisagée et même affinée dans de bonnes conditions.

Exemple 2 : le cas des microblogs Afin de trouver d'autres sources d'URLs potentiellement intéressantes et d'ajouter une composante plus actuelle et proche des utilisateurs, des microblogs et réseaux sociaux sont parcourus afin d'en extraire des URLs dont le potentiel est ensuite déterminé par FLUX. Trois sources sont envisagées : FriendFeed, identi.ca et Reddit. Ces sites sont traversés par des crawls ciblés visant à collecter le maximum d'utilisateurs. Ensuite, la langue des documents derrière les URLs est analysée.

Les résultats montrent qu'il est possible d'obtenir des URLs pointant vers de nombreuses langues différentes. Des centaines si ce n'est des milliers d'URLs ont été trouvées par ce biais pour des langues relativement rares, démontrant l'utilité du processus. Cependant, même en essayant de se concentrer sur des utilisateurs qui ne sont pas anglophones en principe, la part de sites en anglais reste majoritaire.

Comparaison de sources disponibles Les résultats ci-dessus ont ensuite été élargis en termes quantitatifs et linguistiques et comparés avec d'autres sources, notamment des URLs fournies par les moteurs de recherche. Des métrique supplémentaires ont été implémentées, comme la taille des pages, leur âge, leur degré d'interconnexion en termes de liens, les IPs des serveurs hébergeant les sites, ainsi qu'une estimation de la taille de corpus potentielle.

Les résultats ne distinguent aucune source placée clairement au-dessus des autres en termes d'intérêt et de qualité potentiels, démontrant ainsi qu'il est productif d'utiliser plusieurs sources différentes afin de combiner les avantages de chacune d'entre elles. Les résultats de l'étude montrent également qu'il est possible de complémenter si ce n'est de remplacer la méthode BootCaT, tout en soulignant que les moteurs de recherche ne sont pas dominants dans l'état de la recherche par hasard. En effet, ils permettent d'obtenir des pages web au contenu adéquat en termes de longueur ou de langue par exemple. Cependant, le filtrage qu'ils opèrent en termes de contenu est aussi un problème en tant qu'il nuit à la diversité des sites recuellis.

Conclusion Au total, des dizaines de millions d'URLs ont été analysées et annotées par FLUX selon les critères suivants au minimum : langue et intervalle de confiance du détecteur, longueur avant et après le débalisage, nombre de tokens, nombre de liens inter-domaines entrants et sortants, adresse(s) IP du serveur et âge. Si les valeurs moyennes sont encourageantes, les parties basses et surtout hautes des distributions en termes quantitatifs présentent des caractéristiques typiques du web, à savoir un relatif manque de diversité du point de vue du contenu comme des serveurs concernés et des pages sans intérêt destinées à améliorer la crédibilité aux yeux des machines. Qui plus est, l'âge relativement jeune du contenu (moins de six mois en moyenne) incite à reparcourir régulièrement le web pour actualiser les informations.

Impact de la pré-qualification sur le web crawling et la construction de corpus Des expériences sont conduites pour déterminer dans quelle mesure la pré-qualification donne de meilleurs résultats en termes concrets de construction de corpus. À cette fin, le calcul d'un ratio de productivité est exposé, et des crawls sont lancées à partir de sources livrées par FLUX pour différentes sources et différentes langues. Les résultats des procédures de nettoyage des corpus montrent que les sources pré-qualifiées livrent plus de texte supposément de bonne qualité, puisque moins de texte est rejeté lors du pré-traitement. En ce sens, la pré-qualification permet non seulement de donner une direction plus production au crawl mais aussi de gagner du temps, car moins de documents sont téléchargés pour un résultat comparable en quantité.

Une autre expérience à base d'URLs livrées par FLUX montre que les pages web pointées par les URL sont de qualité variable, cependant cette variabilité est surtout fonction de la langue cible et non du type de source.

Téléchargement restreint Le téléchargement restreint implique de déjà connaître la source des documents et d'avoir balisé le terrain afin de ne télécharger que la partie la plus pertinente d'un site web par exemple. Trois exemples sont donnés : des discours politiques, des articles de journaux et des sous-titres de films et séries.

Le cas particulier de blogs sous license CC Un corpus particulier est présenté, il est constitué de blogs en allemand publiés sous licence Creative Commons et à ce titre republiables. Pour le construire, il faut tout d'abord trouver des blogs en allemand, ensuite les parcourir automatiquement puis manuellement pour déterminer s'ils sont publiés sous certaines conditions, et ensuite télécharger le contenu sur de multiples sites. L'expérience montre qu'il faut utiliser les listes d'URLs les plus grandes possibles, et donc éventuellement utiliser des méthodes de web crawling et de pré-qualification afin de pouvoir trouver ces ressources relativement rares.

Qualification de documents et construction de corpus

Corpus généraux Une hiérarchie dans la qualité des documents pour les corpus généraux est définie, des pages web brutes aux textes cohérents et bien-formés. Ensuite, les conditions pratiques des décisions prises lors de la construction de corpus web sont décrites.

1000 textes en anglais pris aux hasard dans les résultats d'un crawl, après pré-traitement, sont manuellement annotés quant à leur adéquation en vue de l'inclusion dans un corpus. Les résultats trois annotateurs, dont deux experts, ne sont que moyennement concordants, malgré une phase de test et d'harmonisation des pratiques. Ce résultat montre qu'il peut être difficile de prendre des décisions quant à l'inclusion ou au rejet de textes tirés du web. En outre, les textes pris en début de crawl s'avèrent être de meilleure qualité que ceux pris en fin de crawl. Ainsi, les questions liées au déroulement d'un crawl ne sont pas triviales et devraient être connues des linguistes et prises en compte.

Détermination de la qualité des textes La qualité du contenu tiré du web ne varie pas seulement d'une page à l'autre mais aussi au fil des documents. En utilisant les méta-données livrées par FLUX, il est possible de jeter un premier regard sur la distribution de certains phénomènes tels que la longueur des documents ou l'importance du marquage.

Selon l'état de la recherche, fixer des seuils de qualité en fonction de mots fréquents ou linguistiquement pertinents, comme les mots du discours, permet d'approcher la question de la qualité de textes tirés du web. Afin d'améliorer ces critères, une batterie de caractéristiques est extraite de l'échantillon manuellement annoté décrit ci-dessus et des méthodes statistiques sont utilisés pour évaluer l'impact des différents critères.

L'étude des corrélations montre que certains critères ont l'air d'apporter plus d'information que d'autres. Des régressions logistiques permettent de mieux appréhender le phénomène. Une combinaison finale d'indicateurs est retenue, elle comprend des indicateurs globaux (ponctuation et espacement), le nombre moyen de tokens par ligne, la proportion de marqueurs du discours au fil du texte, et enfin la proportion de majuscules et de la lettre t. Des comparaisons avec des résultats existants et des validations croisées montrent que cette méthode améliore les résultats de façon consistante, sans pour autant avoir de réelle faiblesse.

Corpus spécialisés Deux corpus spécialisés sont présentés : sous-titres de films et de séries, blogs. Les contraintes et modalités pratiques de construction sont discutées. Les techniques utilisées pour évaluer la qualité des corpus sont décrites : analyses de séries de tokens, recours à des outils d'identification de la langue, utilisation d'une chaîne d'annotation automatique et de statistique afférentes.

Exploitation de corpus

Intérêt et utilisation de corpus Plusieurs corpus spécialisés construits pendant la durée de la thèse sont présentés en termes d'attentes et d'utilisation : presse, discours politiques, sous-titres.

Interfaces Des interfaces pour parcourir et/ou interagir avec les corpus sont présentées, comme par exemple sélection de mots-clés. Des méthodes de comparaison des corpus sont détaillées et un cas d'étude est décrit permettant de comparer des corpus web avec des cor-

pus de références afin premièrement d'évaluer leur qualité et leur cohérence et deuxièmement d'estimer leur apport en ce qui concerne les structures lexicales.

Enfin, les travaux sur la visualisation de corpus sont abordés. Ils résident dans l'extraction de caractéristiques à l'échelle d'un corpus afin de donner des indications sur sa composition et sa qualité.

Conclusion

En conclusion, on peut dire que le web est une chance pour la linguistique et le développement de nouvelles ressources. Cela dit, l'apparence de nouveaux instruments et donc de nouveaux observables suscite de nombreuses questions d'ordre méthodologique. La définition d'une discipline et la scientificité des processus à l'œuvre dans la conception de corpus renvoie aux paradigmes de mégadonnées et de linguistique empirique.

En effet, les corpus parfois très grands tirés du web ne gagnent pas en neutralité à mesure qu'ils grossissent. Un corpus reste reste peu importe sa taille un objet construit, qui est lié à un ensemble de décisions prises. De même, les outils ne peuvent pas être considérés comme étant objectifs, ils appartiennent à un univers technologique et conceptuel qui a un impact sur leur champ d'application et sur leurs résultats.

En ce sens, l'impact des premières étapes de la construction de corpus est sous-évalué, il est important que les linguistes comprennent et puissent conduire les processus de construction et faire évoluer un ensemble de paramètres en fonction de leurs besoins. Les données doivent être soignées et les opérations d'élagage de corpus documentées, afin de montrer que les corpus web ne sont pas des objets jetables mais bien des objets scientifiques qui impliquent un véritable travail.

Les acquis de la science du web devraient être mis à profit pour construire des corpus. Ainsi, les approches légères peuvent prouver leur efficacité et être adaptées à l'envi avant l'emploi de méthodes plus gourmandes en temps et en ressources. La création d'un annuaire d'URLs préqualifiées s'inscrit dans cette démarche, qui construit non seulement des corpus mais également des outils et des bases de données évolutives et donc en phase avec les changement du web.

Enfin, il y a un lien qui ne devrait pas être ignoré entre l'art des jardins, la culture et la connaissance. Le processus de découverte scientifique peut tout à fait surgir de la joie de disposer d'un cabinet de curiosités ainsi que d'une recherche esthétique dans les visualisations, afin de parler à l'œil autant qu'à la raison.

References

Abney, S. (1996). Statistical methods and linguistics. In *The balancing act: Combining symbolic and statistical approaches to language* (pp. 1–26). MIT Press.

Abney, S., & Bird, S. (2010). The Human Language Project: building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the ACL* (pp. 88–97). Association for Computational Linguistics.

Abney, S. P. (1991). Parsing by chunks. *Principle-based parsing*, *44*, 257–278.

Abramson, M., & Aha, D. W. (2012). What's in a URL? Genre Classification from URLs. In *Intelligent techniques for web personalization and recommender systems. aaai technical report.* Association for the Advancement of Artificial Intelligence.

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, *30*(2), 141–161.

Arase, Y., & Zhou, M. (2013). Machine Translation Detection from Monolingual Web-Text. In *Proceedings of the 51th Annual Meeting of the ACL* (pp. 1597–1607).

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1–16.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Auroux, S. (1998). *La raison, le langage et les normes.* PUF.

Austin, P. K. (2010). Current issues in language documentation. *Language documentation and description*, *7*, 12–33.

Bachelard, G. (1927). *Essai sur la connaissance approchée.* Vrin.

Bader, M., & Häussler, J. (2010). Word Order in German: A Corpus Study. *Lingua*, *120*(3), 717–762.

Baisa, V. (2009). *Web content cleaning.* Unpublished master's thesis, Faculty of Informatics, Masaryk university.

Baker, P., Hardie, A., McEnery, T., Xiao, R., Bontcheva, K., Cunningham, H., et al. (2004). Corpus linguistics and South Asian languages: Corpus creation and tool development. *Literary and Linguistic Computing*, *19*(4), 509–524.

Barbaresi, A. (n.d.). *German Political Speeches, Corpus and Visualization* (Tech. Rep.). ICAR / ENS Lyon. (2nd Version, presented at the DGfS-CL poster session)

Barbaresi, A. (2011a). Approximation de la complexité perçue, méthode d'analyse. In *Actes*

TALN'2011/RECITAL (Vol. 2, pp. 229–234). Montpellier, France.

Barbaresi, A. (2011b). *Théorie, terrain et techniques : quels recoupements et quelles médiations ?* (Tech. Rep.). ENS Lyon. (Rencontre méthodologique ENthèSe, Pour une nouvelle appréhension des textes et données textuelles : pratiques, outils, méthodes)

Barbaresi, A. (2012). La Raison aveugle ? L'époque cybernétique et ses dispositifs. In *Les critiques de la raison au XXe siècle.* Université Paris-Est Créteil, France.

Barbaresi, A. (2013a). Challenges in web corpus construction for low-resource languages in a post-BootCaT world. In Z. Vetulani & H. Uszkoreit (Eds.), *Proceedings of the 6th language & technology conference, less resourced languages special track* (pp. 69–73). Poznań.

Barbaresi, A. (2013b). Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop* (pp. 9–15).

Barbaresi, A. (2014a). Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. In R. Schäfer & F. Bildhauer (Eds.), *Proceedings of the 9th web as corpus workshop* (pp. 1–8).

Barbaresi, A. (2014b). *Language-classified Open Subtitles (LACLOS): Download, extraction, and quality assessment* (Tech. Rep.). BBAW.

Barbaresi, A., & Würzner, K.-M. (2014). For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *KONVENS 2014, NLP4CMC workshop proceedings* (pp. 2–10). Hildesheim University Press.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC* (pp. 1313–1316).

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Baroni, M., Chantree, F., Kilgarriff, A., & Sharoff, S. (2008). Cleaneval: a Competition for Cleaning Web Pages. In *Proceedings of lrec.*

Baroni, M., & Ueyama, M. (2006). Building general- and special-purpose corpora by Web crawling. In *Proceedings of the 13th nijl international symposium, language corpora: Their compilation and application* (pp. 31–40).

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699–725.

Baykan, E., Henzinger, M., Marian, L., & Weber, I. (2009). Purely URL-based topic classification. In *Proceedings of the 18th international conference on world wide web* (pp. 1109–1110).

Baykan, E., Henzinger, M., & Weber, I. (2008). Web Page Language Identification Based on URLs. *Proceedings of the VLDB Endowment*, *1*(1), 176–187.

Beißwenger, M. (2007). Corpora zur computervermittelten (internetbasierten) Kommunikation. *Zeitschrift für germanistische Linguistik*, *35*(3), 496–503.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., & Storrer, A. (2013). DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, *28*(4), 531–537.

Bendersky, M., Croft, B. W., & Diao, Y. (2011). Quality-biased ranking of web documents. In *Proceedings of the 4th ACM conference on Web search and data mining* (pp. 95–104).

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on Twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (ceas)*

(Vol. 6, pp. 12–21).
Bensaude-Vincent, B. (2009). *Les vertiges de la technoscience.* Paris: La Découverte.
Bergh, G., & Zanchetta, E. (2008). Web linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics, An International Handbook* (pp. 309–327). De Gruyter.
Bergounioux, G. (1992). Les enquêtes de terrain en France. *Langue française*, *93*, 3–22.
Bergsma, S., Lin, D., & Goebel, R. (2009). Web-Scale N-gram Models for Lexical Disambiguation. In *IJCAI* (Vol. 9, pp. 1507–1512).
Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., & Weitzner, D. J. (2006). A Framework for Web Science. *Foundations and Trends in Web Science*, *1*(1), 1–130.
Berthelot, J.-M. (Ed.). (2003). *Figures du texte scientifique.* PUF.
Bertin, J. (1967). *Sémiologie graphique : les diagrammes, les réseaux, les cartes.* Paris: Bordas.
Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, *15*(2), 133–163.
Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, *8*(4), 243–257.
Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge University Press.
Biber, D., Conrad, S., & Reppen, R. (2006). *Corpus linguistics - Investigating language structure and use* (5th ed.). Cambridge: Cambridge University Press.
Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: a multidimensional analysis. In M. Hundt, N. Nadja, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 109–127). Rodopi.
Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R., et al. (2013). Scalable Construction of High-Quality Web Corpora. *Journal for Language Technology and Computational Linguistics*, 23–59.
Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig corpora collection – Monolingual corpora of standard size. In *Proceedings of Corpus Linguistics.*
Blache, P. (2011). Evaluating Language Complexity in Context: New Parameters for a Constraint-Based Model. In P. Blache, H. Christiansen, V. Dahl, & J. Villadsen (Eds.), *Constraints and Language Processing, 6th International Workshop, CSLP 2011* (pp. 7–20).
Blanchet, P. (2003). Contacts, continuum, hétérogénéité, polynomie, organisation chaotique, pratiques sociales, interventions… quels modèles? Pour une (socio) linguistique de la complexité. *Langues, contacts, complexité. Perspectives théoriques en sociolinguistique*, 279–308.
Borin, L. (2009a). Linguistic diversity in the information society. In *Proceedings of the saltmil 2009 workshop on information retrieval and information extraction for less resourced languages* (pp. 1–7).
Borin, L. (2009b). *One in the bush. Low-density language technology* (Tech. Rep.). University of Gothenburg.
Borin, L., Dubhashi, D., Forsberg, M., Johansson, R., Kokkinakis, D., & Nugues, P. (2013). Mining semantics for culturomics: towards a knowledge-based approach. In *Proceedings of the international workshop on mining unstructured big data using natural language processing* (pp. 3–10).
Boulton, A. (2013). Wanted: Large corpus, simple software. No timewasters. In *TaLC10: Proceedings of the 10th International Conference on Teaching and Language Corpora*

(pp. 1–6).

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Brants, T., & Franz, A. (2006). *Web 1t 5-gram version 1* (Tech. Rep.). Linguistic Data Consortium.

Bretan, I., Dewe, J., Hallberg, A., Wolkert, N., & Karlgren, J. (1998). Web-Specific Genre Visualization. In *Webnet*.

Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G. (1997). Syntactic clustering of the web. *Computer Networks and ISDN Systems*, *29*(8), 1157–1166.

Bronckart, J.-P., Bain, D., Schneuwly, B., Davaud, C., & Pasquier, A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*. Lausanne: Delachaux & Niestlé.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Bungarten, T. (1979). Das korpus als empirische grundlage in der linguistik und literaturwissenschaft. *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. Monografien Linguistik und Kommunikationswissenschaft*, *39*.

Burbulla, J. (2011). Schlachtrufe großer Geister. Gartenkunst als nationale Versuchs=Kunst der Frühen Neuzeit. In J. Burbulla & A.-S. Tabarasi-Hoffmann (Eds.), *Gartenkunst und Wissenschaft : Diskurs, Repräsentation, Transformation seit dem Beginn der Frühmoderne* (pp. 9–48). Peter Lang.

Burnard (ed.), L. (2007). *Reference Guide for the British National Corpus (XML Edition)* (Tech. Rep.). Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services. (http://www.natcorp.ox.ac.uk/docs/URG/index.html)

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies ased on film subtitles. *PLoS One*, *5*(6), e10729.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, *22*(2), 249–254.

Celli, F. (2009). *Improving Language identification performance with FriendFeed data* (Tech. Rep.). CLIC, University of Trento.

Chater, N., & Christiansen, M. H. (2008). Computational Models of Psycholinguistics. In R. Sun (Ed.), *The Cambridge Handbook of Psycholinguistics*. Cambridge University Press.

Chevalier, J.-C. (1997). Le baron de Tourtoulon et la constitution d'une géographie linguistique. *Lengas*, *21*(42), 163–170.

Chow, S., & Ruskey, F. (2004). Drawing area-proportional venn and euler diagrams. In G. Liotta (Ed.), *Graph drawing* (Vol. 2912, pp. 466–477). Springer.

Church, K. W., & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational linguistics*, *19*(1), 1–24.

Clément, G. (1991). *Le Jardin en mouvement*. Pandora.

Collins, C., Viegas, F. B., & Wattenberg, M. (2009). Parallel tag clouds to explore and analyze faceted text corpora. In *Visual analytics science and technology* (pp. 91–98).

Cori, M. (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. *Langages*, *171*(3), 95–110.

Cox, R. (2007). *Regular Expression Matching Can Be Simple And Fast (but is slow in Java, Perl, PHP, Python, Ruby, ...).* (http://swtch.com/ rsc/regexp/regexp1.html)

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213–238.

Crawford, K., Gray, M., & Miltner, K. (2014). Big Data | Critiquing Big Data: Politics, Ethics, Epistemology | Special Section Introduction. *International Journal of Communication*, *8*, 1663–1672. Available from http://ijoc.org/index.php/ijoc/article/view/2167

Crowgey, J., & Bender, E. M. (2011). Analyzing Interacting Phenomena: Word Order and Negation in Basque. In S. Müller (Ed.), *Proceedings of the hpsg11 conference.* CSLI Publications.

Daoust, F., Laroche, L., & Ouellet, L. (1996). Sato-Calibrage: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, *25*(1), 205–234.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011). READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies* (pp. 73–83). Edinburgh, Scotland, UK: Association for Computational Linguistics.

De Schryver, G.-M. (2002). Web for/as corpus: A perspective for the African languages. *Nordic Journal of African Studies*, *11*(2), 266–282.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Durand, J. (2009). On the scope of linguistics: data, intuitions, corpora. In Y. Kawaguchi, M. Minegishi, & J. Durand (Eds.), *Corpus and variation in linguistic description and language education* (pp. 25–52). John Benjamins.

Dürscheid, C. (2003). Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. *Zeitschrift für Angewandte Linguistik*, *38*, 35–54.

Eckart, T., Quasthoff, U., & Goldhahn, D. (2012). The Influence of Corpus Quality on Statistical Measurements on Language Resources. In *Proceedings of lrec* (pp. 2318–2321).

Eichinger, L. (2006). Linguisten brauchen Korpora und Korpora Linguisten. In *Sprachkorpora - Datenmengen und Erkenntnisfortschritt (Institut für Deutsche Sprache – Jahrbuch 2006)* (pp. 1–7). De Gruyter.

Evert, S. (2008). A Lightweight and Efficient Tool for Cleaning Web Pages. In *Proceedings of lrec.*

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 276–284).

Fillmore, C. J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". In J. Svartvik (Ed.), *Directions in Corpus Linguistics* (pp. 35–60). Berlin, New York: Mouton de Gruyter.

Fischer, H., Remmert, V., & Wolschke-Bulmahn, J. (2011). Wissen und Gärten. Gartenkunst und Naturwissenschaften in der Frühen Neuzeit. Mathematisierung und Verwissenschaftlichung in der frühneuzeitlichen Gartenkunst. In J. Burbulla & A.-S. Tabarasi-Hoffmann (Eds.), *Gartenkunst und Wissenschaft : Diskurs, Repräsentation, Transformation seit dem Beginn der Frühmoderne* (pp. 271–292). Peter Lang.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2005). Visualization of text document corpus. *Informatica*, *29*(4).

François, T. (2009). Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE. In *Actes TALN/RECITAL*. Senlis: ATALA.

François, T., & Fairon, C. (2012, July). An "AI readability" Formula for French as a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 466–477). Jeju Island, Korea: Association for Computational Linguistics.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, *31*(7), 931 - 952.

Freywald, U. (2009). Kontexte für nicht-kanonische Verbzweitstellung: V2 nach dass und Verwandtes. In V. Ehrich, C. Fortmann, I. Reich, & M. Reis (Eds.), *Koordination und subordination im deutschen* (pp. 113–134). Buske.

Gerdes, K. (2014). Corpus collection and analysis for the linguistic layman: The Gromoteur. In *Proceedings of JADT 2014* (pp. 261–).

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects* (pp. 23–41). Continuum Press.

Geyken, A., & Hanneforth, T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing* (Vol. 4002, pp. 55–66). Springer.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A Noisy-Channel Account of Crosslinguistic Word-Order Variation. *Psychological Science*, 1079–1088.

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2011). Practical recommendations on crawling online social networks. *IEEE Journal on Selected Areas in Communications*, *29*(9), 1872–1892.

Glazer, S. (1974). Is Sentence Length a Valid Measure of Difficulty in Readability Formulas? *The Reading Teacher*, *27*(5), 464–468.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of LREC* (pp. 759–765).

Grabe, W., & Stoller, F. L. (2013). *Teaching and researching: Reading*. Routledge.

Graën, J., Batinic, D., & Volk, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. In *Proceedings of KONVENS 2014* (pp. 222–227).

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241.

Grieve, J., Asnaghi, C., & Ruette, T. (2013). Site-restricted web searches for data collection in regional dialectology. *American Speech*, *88*(4), 413–440.

Gupta, T., Garg, S., Carlsson, N., Mahanti, A., & Arlitt, M. (2009). Characterization of FriendFeed – A Web-based Social Aggregation Service. In *Proceedings of the aaai icwsm* (Vol. 9).

Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment. In M. Bilger (Ed.), *Linguistique sur corpus. Etudes et réflexions* (p. 11-58).

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus*. Armand Colin.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *Intelligent Systems*, *24*(2), 8–12.

Halliday, M. A. K. (1992). Language as system and language as instance: The corpus as a

theoretical construct. In J. Svartvik (Ed.), *Directions in corpus linguistics. proceedings of the nobel symposium 82* (Vol. 65, pp. 61–77). Berlin/New York: Mouton de Gruyter.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Heilman, M., Collins-Thompson, K., & Eskenazi, M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the third workshop on innovative use of nlp for building educational applications* (pp. 71–79).

Heister, J., & Kliegl, R. (2012). Comparing word frequencies from different German text corpora. In K.-M. Würzner & E. Pohl (Eds.), *Lexical Resources in Psycholinguistic Research* (pp. 27–44). Potsdam Cognitive Science Series. (vol.3)

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*(62), 10–20.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., & Weitzner, D. (2008). Web Science: An Interdisciplinary Approach to Understanding the Web. *Communications of the ACM*, *51*(7), 60–69.

Henzinger, M. (2006). Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 284–291).

Heyer, G., & Quasthoff, U. (2004). Calculating Communities by Link Analysis of URLs. In *Proceedings of IICS-04, Guadalajara, Mexico* (Vol. 3473, pp. 151–156). Berlin/Heidelberg: Springer.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for. *Essentials of language documentation*, *178*, 1–30.

Hinrichs, E. W. (2005). Finite-State Parsing of German. In A. Arppe & et al. (Eds.), *Inquiries into Words, Constraints and Contexts* (p. 35-44). Stanford: CSLI Publications.

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., et al. (1997). FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *Finite-State Language Processing*, 383–406.

Hoffmann, S. (2006). From web page to mega-corpus: the cnn transcripts. *Language and Computers*, *59*(1), 69–85.

Hottois, G. (1984). *Le signe et la technique : la philosophie a` l'e´preuve de la technique*. Paris: Aubier.

Hundt, M. (2008). Text corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics, An International Handbook* (pp. 168–187). De Gruyter.

Hundt, M., Nesselhauf, N., & Biewer, C. (2007). Corpus linguistics and the web, introduction. In M. Hundt, N. Nadja, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 1–5). Rodopi.

Hunston, S. (2008). Collection strategies and design decisions. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics, An International Handbook* (pp. 154–168). De Gruyter.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., Suchomel, V., et al. (2013). The TenTen Corpus Family. In *Proceedings of the international conference on corpus linguistics*.

Jessop, M. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing*, *23*(3), 281–293.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Johansson, S. (2008). Some aspects of the development of corpus linguistics in the 1970s and 1980s. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics, An International Handbook* (pp. 33–53). De Gruyter.

Joseph, B. D. (2004). The Editor's Department: On change in Language and change in language. *Language*, *80*(3), 381–383.

Jurish, B. (2003). *A Hybrid Approach to Part-of-Speech Tagging* (Final Report). Kollokationen im Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften.

Jurish, B., & Würzner, K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, *28*(2), 61-83. Available from `http://www.jlcl.org/2013_Heft2/3Jurish.pdf`

Karmakar, S., & Zhu, Y. (2011). Visualizing Text Readability. In *6th international conference on advanced information management and service (ims)* (pp. 291–296).

Karttunen, L. (2001). Applications of Finite-State Transducers in Natural Language Processing. In S. Yu & A. Paun (Eds.), *Ciaa 2000, lncs 2088* (pp. 34–46). Heidelberg: Springer.

Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., et al. (2010). Learning to Predict Readability using Diverse Linguistic Features. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 546–554).

Kay, M. (2005). A Life of Language. *Computational Linguistics*, *31*(4), 425–438.

Kermes, H., & Evert, S. (2002). YAC – A Recursive Chunker for Unrestricted German Text. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (Vol. 5, pp. 1805–1812).

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650.

Kilgarriff, A. (2007). Googleology is bad science. *Computational linguistics*, *33*(1), 147–151.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, *29*(3), 333–347.

Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. (2010). A Corpus Factory for Many Languages. In *Proceedings of LREC* (pp. 904–910).

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. *Information Technology*, *105*, 116–127.

King, B., & Abney, S. (2013). Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of naacl-hlt* (pp. 1110–1119).

Kintsch, W., & Van Dijk, T. A. (1978). Toward a Model of Text Comprehension and Production. *Psychological Review*, *85*(5), 363–394.

Klebanov, B. B., & Flor, M. (2013). Word Association Profiles and their Use for Automated Scoring of Essays. In *Proceedings of the 51st Annual Meeting of the ACL* (pp. 1148–1158).

Kohlschütter, C., Fankhauser, P., & Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the 3rd acm international conference on web search and data mining* (pp. 441–450). New York: ACM.

Koster, M. (1994). *A standard for robot exclusion* (Tech. Rep.). (http://www.robotstxt.org/orig.html)

Kreyer, R. (2006). *Inversion in Modern Writtern English, Syntactic Complexity, Information Status and the Creative Writer*. Tübingen: Günter Narr.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Beverly Hills: Sage Publications.

Latour, B. (1985). Les "vues" de l'esprit. *Culture technique*, *14*, 4–29.

Leech, G. (2006). New resources, or just better old ones? The Holy Grail of representativeness. *Language and Computers*, *59*(1), 133–149.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, *16*(9).

Lehmann, C. (2006). Daten – Korpora – Dokumentation. In *Sprachkorpora - Datenmengen und Erkenntnisfortschritt (Institut für Deutsche Sprache – Jahrbuch 2006)* (pp. 9–27). De Gruyter.

Lejeune, G. (2013). *Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel.* Unpublished doctoral dissertation, University of Caen.

Lemnitzer, L., & Zinsmeister, H. (2010). *Korpuslinguistik: Eine Einführung* (2nd ed.). Gunter Narr.

Léon, J. (2005). Claimed and unclaimed sources of corpus linguistics. *Henry Sweet Society Bulletin*, *44*, 36–50.

Loiseau, S. (2008). Corpus, quantification et typologie textuelle. *Syntaxe et sémantique*, *9*, 73–85. Available from http://hal.archives-ouvertes.fr/halshs-00374645/

Louis, A., & Nenkova, A. (2012). A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1157–1168).

Lui, M., & Baldwin, T. (2011). Cross-domain Feature Selection for Language Identification. In *Proceedings of the fifth international joint conference on natural language processing (ijcnlp 2011)* (pp. 553–561). Chiang Mai, Thailand. Available from http://www.aclweb.org/anthology/I11-1062

Lui, M., & Baldwin, T. (2012). langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th annual meeting of the association for computational linguistics (acl 2012).* Jeju, Republic of Korea.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., et al. (2014). The paisa corpus of italian web texts. *Proceedings of the 9th Web as Corpus Workshop*, 36–43.

Lüdeling, A. (2006). Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In *Sprachkorpora - Datenmengen und Erkenntnisfortschritt (Institut für Deutsche Sprache – Jahrbuch 2006)* (pp. 28–48). De Gruyter.

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nadja, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 7–24). Rodopi.

Mair, C. (2012). From opportunistic to systematic use of the web as corpus. *The Oxford Handbook of the History of English*, 245–255.

Marshall, N., & Glock, M. D. (1978). Comprehension of Connected Discourse: A Study into the Relationships between the Structure of Text and Information Recalled. *Reading Research Quarterly*, *14*(1), 10–56.

McCarthy, M., & O'Keeffe, A. (2010). What are corpora and how have they evolved? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 3–13). Routledge.

McEnery, T. (2003). Corpus Linguistics. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 448–463). Oxford Universtity Press.

McNamara, D., Louwerse, M., McCarthy, P., & Graesser, A. (2010). Coh-metrix: Capturing

linguistic features of cohesion. *Discourse Processes*, *47*(4), 292–330.

Meyer, C. F. (2008). Pre-electronic corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics, An International Handbook* (pp. 1–14). De Gruyter.

Mohr, G., Stack, M., Rnitovic, I., Avery, D., & Kimpton, M. (2004). Introduction to Heritrix. In *Proceedings of the 4th international web archiving workshop (iwaw'04).*

Müller, F. H. (2007). *A Finite-State Approach to Shallow Parsing and Grammatical Functions Annotation of German.* Unpublished doctoral dissertation, University of Tübingen.

Neumann, G., Backofen, R., Baur, J., Becker, M., & Braun, C. (1997). An Information Extraction Core System for Real World German Text Processing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (pp. 209–216).

Oelke, D., Spretke, D., Stoffel, A., & Keim, D. A. (2010). Visual readability analysis: How to make your writings easier to read. In *VAST 10: IEEE Conference on Visual Analytics Science and Technology* (pp. 123–130).

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge Handbook of Corpus Linguistics* (1st ed.). London, New York: Routledge.

Olston, C., & Najork, M. (2010). Web Crawling. *Foundations and Trends in Information Retrieval*, *4*(3), 175–246.

A one-pass valency-oriented chunker for German. (n.d.).

Ostler, N. (2008). Corpora of less studied languages. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics, An International Handbook* (pp. 457–484). De Gruyter.

Ott, N. (2009). *Information Retrieval for Language Learning: An Exploration of Text Difficulty Measures.* Unpublished master's thesis, University of Tübingen, Seminar für Sprachwissenschaft, Tübingen, Germany.

Pereira, F. (1990). Finite-state approximations of grammars. In *Proceedings of the annual meeting of the acl* (pp. 20–25).

Perkuhn, R., Keibel, H., & Kupietz, M. (2012). *Korpuslinguistik.* Wilhelm Fink.

Pincemin, B. (2006). Introduction. *Corpus*, *6*, 5–15.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 186–195).

Préfontaine, C., & Lecavalier, J. (1996). Analyse de l'intelligibilité de textes prescriptifs. *Revue québécoise de linguistique*, *25*(1), 99–144.

R Core Team. (2012a). *R: A language and environment for statistical computing* (Tech. Rep.). (http://www.R-project.org/)

R Core Team. (2012b). R: A language and environment for statistical computing [Computer software manual]. (http://www.R-project.org/)

Rarrick, S., Quirk, C., & Lewis, W. (2011). MT Detection in Web-Scraped Parallel Corpora. In *Proceedings of the Machine Translation Summit (MT Summit XIII).*

Rayson, P., & Mariani, J. (2009). Visualising corpus linguistics. In *Proceedings of the corpus linguistics conference.*

Rehm, G., Witt, A., Zinsmeister, H., & Dellert, J. (2007). Corpus masking: Legally bypassing licensing restrictions for the free distribution of text collections. *Digital Humanities*, 166–170.

Reis, M. (1980). On justifying topological frames: 'Positional field' and the order of nonverbal constituents in German. *DRLAV*, *22*(23), 59–85.

Renouf, A. (1993). A Word in Time: first findings from the investigation of dynamic text. In

English Language Corpora: Design, Analysis and Exploitation (pp. 279–288). Rodopi.

Renouf, A. (2007). Corpus development 25 years on: from super-corpus to cyber-corpus. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on.* Rodopi.

Resnik, P., & Elkiss, A. (2005). The linguist's search engine: an overview. In *Proceedings of the annual meeting of the acl* (pp. 33–36).

Resnik, P., & Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, *29*(3), 349–380.

Riloff, E., & Phillips, W. (2004). *An Introduction to the Sundance and AutoSlog Systems* (Tech. Rep.). School of Computing, University of Utah.

Rohrer, R. M., Ebert, D. S., & Sibert, J. L. (1998). The shape of shakespeare: Visualizing text using implicit surfaces. In *Proceedings of the ieee symposium on information visualization* (pp. 121–129).

Rundell, M. (2008). The corpus revolution revisited. *English Today*, *24*(01), 23–27.

Rundell, M., & Stock, P. (1992). The corpus revolution. *English Today*, *8*(03), 21–32.

Salomon, D., & Motta, G. (2010). *Handbook of Data Compression.* London: Springer.

Sampson, G. (2000). The role of taxonomy in language engineering. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *358*(1769), 1339–1355.

Sampson, G. (2013). The empirical trend – Ten years on. *International Journal of Corpus Linguistics*, *18*(2), 281–289.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop* (Vol. 4, pp. 5–15).

Schäfer, R., Barbaresi, A., & Bildhauer, F. (2013). The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In S. Evert, E. Stemle, & P. Rayson (Eds.), *Proceedings of the 8th web as corpus workshop* (pp. 7–15).

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *Proceedings of LREC* (pp. 486–493).

Schiehlen, M. (2003). A Cascaded Finite-State Parser for German. In *Proceedings of the 10th conference of the eacl* (Vol. 2, pp. 163–166).

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1995). *Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS* (Draft). Universities of Stuttgart and Tübingen.

Schmid, H. (1994). Probabilistic Part-Of-Speech Tagging Using Decision Trees. In *Proceedings of the international conference on new methods in language processing* (Vol. 12).

Schmid, H., & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd international conference on computational linguistics* (Vol. 1, pp. 777–784).

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the acl* (pp. 523–530).

Schäfer, R., Barbaresi, A., & Bildhauer, F. (2014). Focused Web Corpus Crawling. In F. Bildhauer & R. Schäfer (Eds.), *Proceedings of the 9th Web as Corpus workshop (WAC-9)* (pp. 9–15).

Schäfer, R., & Bildhauer, F. (2013). *Web Corpus Construction.* Morgan & Claypool.

Sharoff, S. (2004). Towards basic categories for describing properties of texts in a corpus. In *Proceedings of lrec.*

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries.

In *Wacky!* (pp. 63–98).
Sinclair, J. (1996). *Preliminary recommendations on Corpus Typology* (Tech. Rep.). EAGLES – Expert Advisory Group on Language Engineering Standards.
Sinclair, J. (2008). Borrowed ideas. In Gerbig, Andrea and Mason, Oliver (Ed.), *Language, People, Numbers: Corpus Linguistics and Society* (pp. 21–42). Rodopi.
Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., & Lopez, A. (2013). Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of the 51th annual meeting of the acl* (pp. 1374–1383).
Snow, C. E. (2002). *Reading for Understanding : Toward an R&D Program in Reading Comprehension.* RAND.
Spoustová, J., & Spousta, M. (2012). A High-Quality Web Corpus of Czech. In *Proceedings of LREC* (pp. 311–315).
Steger, J., & Stemle, E. (2009). KrdWrd, Architecture for Unified Processing of Web Content. In *Proceedings of the fifth web as corpus workshop (wac5)* (pp. 63–70).
Storrer, A. (2001). Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation. In A. Lehr, M. Kammerer, K.-P. Konerding, A. Storrer, C. Thimm, & W. Wolski (Eds.), *Sprache im alltag. beiträge zu neuen perspektiven in der linguistik* (pp. 439–466). De Gruyter.
Suchomel, V., & Pomikálek, J. (2012). Efficient Webcrawling for large text corpora. In A. Kilgarriff & S. Sharoff (Eds.), *Proceedings of the 7th Web as Corpus Workshop* (pp. 40–44).
Swinton, J. (2009). *Vennerable.* http://r-forge.r-project.org/projects/vennerable.
Tabarasi-Hoffmann, A.-S. (2011). "In a small Compass, a model of Universal Nature made private". Zur Verbindung von Garten, Labor, Bibliothekund Kunstkammer im England des 17. Jahrhunderts. In J. Burbulla & A.-S. Tabarasi-Hoffmann (Eds.), *Gartenkunst und Wissenschaft : Diskurs, Repräsentation, Transformation seit dem Beginn der Frühmoderne* (pp. 127–162). Peter Lang.
Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting Texts by Readability. *Computational Linguistics*, *36*(2), 203–227.
Tanguy, L. (2012). *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes.* (Habilitation thesis, University of Toulouse II-Le Mirail)
Tanguy, L. (2013). La ruée linguistique vers le Web. *Texto ! Textes et cultures*, *4*(18).
Teubert, W. (2010). Corpus linguistics: An alternative. *Semen. Revue de sémio-linguistique des textes et discours*(27).
The Royal Society Science Policy Center. (2012). *Science as an open enterprise* (Tech. Rep.). The Royal Society.
Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* John Benjamins.
Tognini Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 14–27). Routledge.
Vajjala, S., & Meurers, D. (2013). On The Applicability of Readability Models to Web Texts. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 59–68).
Valette, M. (2008). Pour une science des textes instrumentée. *Syntaxe et sémantique*, *9*, 9–14.
Váradi, T. (2001). The linguistic relevance of corpus linguistics. In *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 587–593).

Venn, J. (1880). On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *10*(59), 1–18.

Voss, M. J. (2005). *Determining syntactic complexity using very shallow parsing.* (Master's thesis, CASPR Research Report, Artificial Intelligence Center, University of Georgia)

Wallis, S., & Nelson, G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, *5*(4), 305–335.

Warschauer, M., & Grimes, D. (2007). Audience, Authorship, and Artifact: The emergent semiotics of Web 2.0. *Annual Review of Applied Linguistics*, *27*, 1–23.

Xiao, R. (2008). Well-known and influential corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics, An International Handbook* (pp. 383–457). De Gruyter.

Index