Web Corpus Creation and Cleaning

egon w. stemle <egon.stemle@eurac.edu>

European Academy of Bozen/Bolzano
(EURAC)

July 13th, 2012

Web corpus creation. Web corpus cleaning.

It seems appropriate to talk about:

- Web corpora (plural of corpus)
- Web corpus creation
- Web corpus cleaning

(Very simple) Ways to exploit a Corpus

EURAC research

- Type and token (frequency) lists
  words (abstract $\mathfrak{go}$ vs. go/went/gone) that appear in a corpus, and their frequencies
- Word N-grams
  a moving window over a text, where the window size is N words
- Concordances, collocations/collegations
  words in the context they appear; occurrence of (specific) words within a pre-defined distance
- Specifically designed programs (especially when the corpus is annotated)

(A few) existing Corpora

EURAC research

- AmE Brown and BrE Lancaster/Oslo/Bergen (LOB) C. ca.1960/80
  1 million (M) word collection of 500 texts of around 2000 words each, distributed across 15 text categories, 9 informative and 6 imaginative
- Wall Street Journal (WSJ) and Reuters newswire C. ca.1987/2000
  30M/1.3M words of news stories
- British National Corpus (BNC) ca.1993
  100M word collection of samples of written and spoken BrE language of the late 20th century, from a wide variety of genres
- De/It/UkWaC Web Corpora 2008
  approx. 2 billion word collections of written German/Italian/English from the web, from an unknown variety of genres
- Paisà Web Corpus 2011
  collection of written Italian from the web, containing only creative commons (CC) licensed text

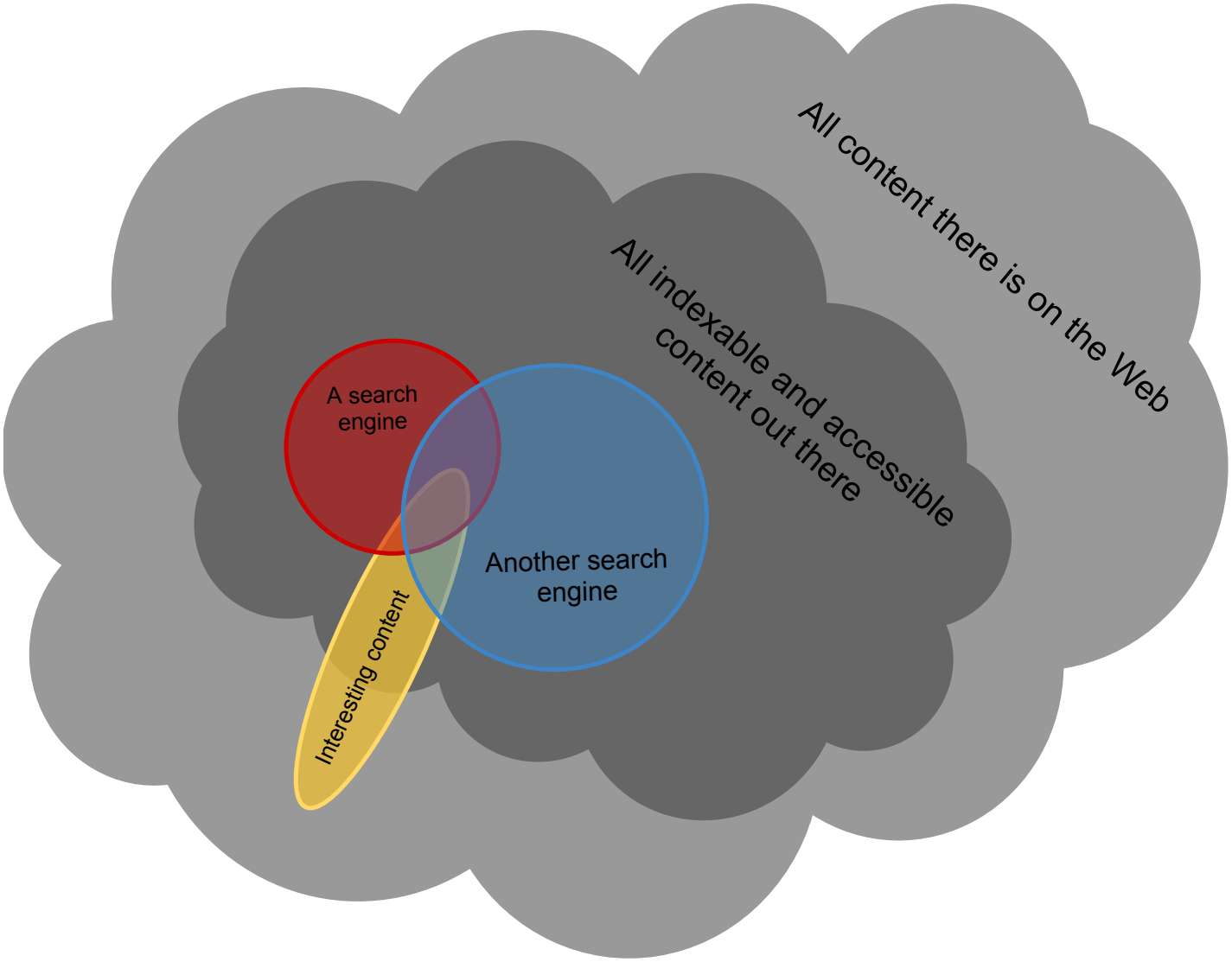
The big Picture

EURAC research

Statement (Assumption)

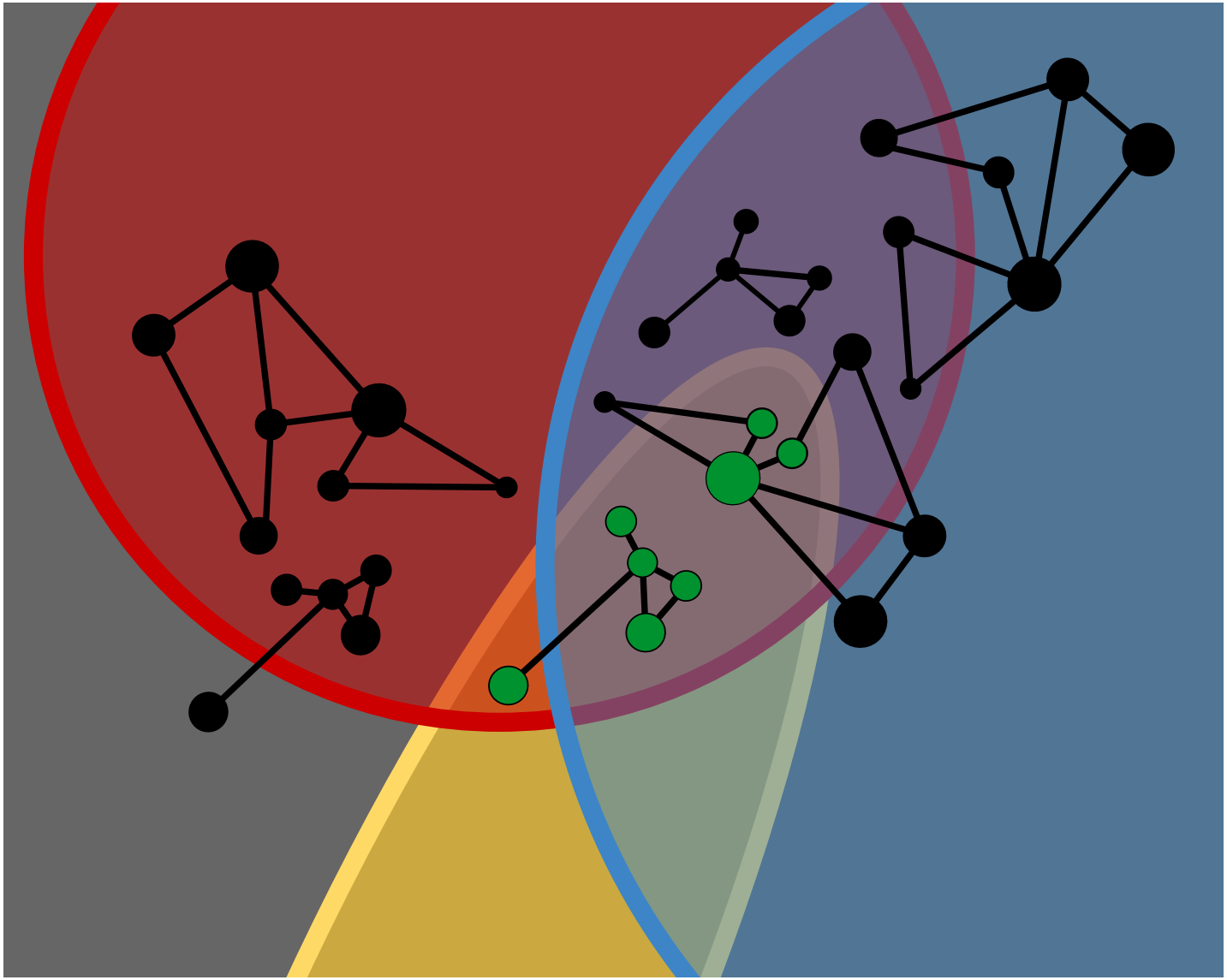
The Web is an unprecedented and virtually inexhaustible source of authentic natural language data and offers the HLT community an opportunity to train statistical models on much larger amounts of data than was previously possible.

The really big Picture



The really big Picture

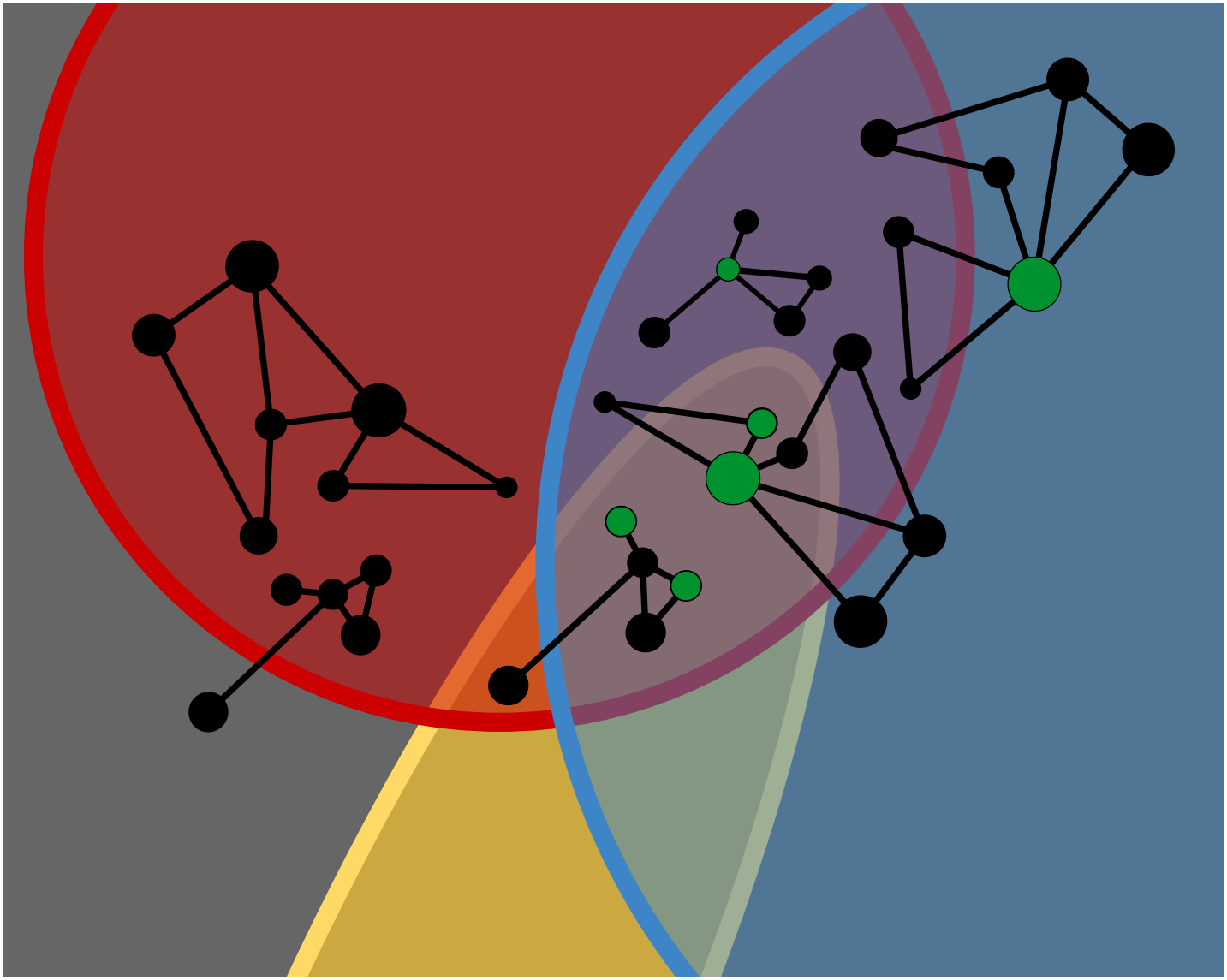
(closeup)



cross-linked structure of web pages and *the* interesting web pages

The really big Picture

(closeup)



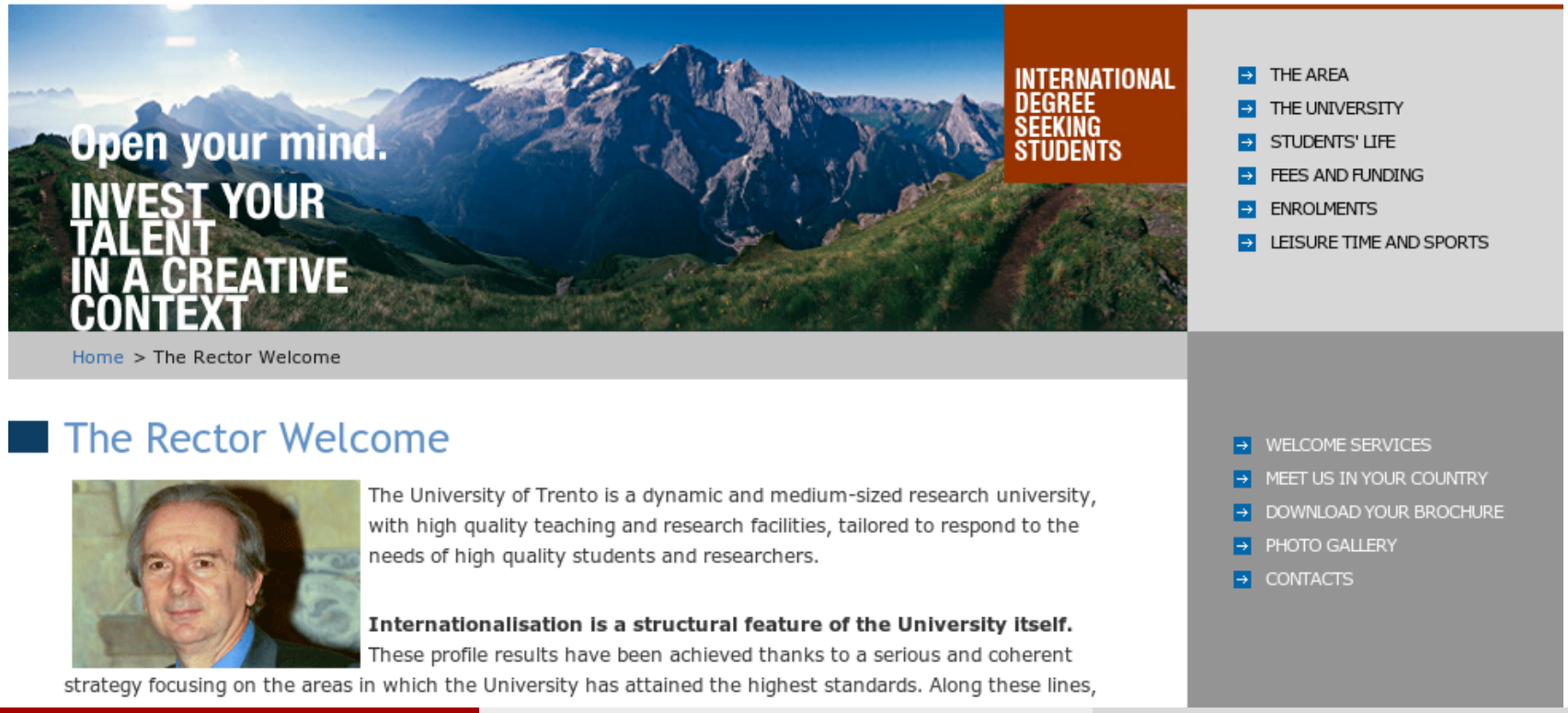
web pages returned by some means of query

(Some) Things can go wrong – and if they can they will. . .

EURAC research

Documents might turn out to

1. be in the wrong language (or of the wrong genre)



(Some) Things can go wrong – and if they can they will. . .

EURAC research

Documents might turn out to

1. be in the wrong language (or of the wrong genre)
2. contain gibberish (characters, words, . . .)

Wikipedia

Wikipedia [ËŒvÉªkiËˆpeËdia] (auch: *die Wikipedia*) ist ein am 15. Januar 2001 gegrÃ¼ndetes freies Onlin
Wikipedia ist ein Kofferwort, das sich aus â€žWikiâ€œ (hawaiisch fÃ¼r â€žschnellâ€œ) und â€žEncyclope
¤die) zusammensetzt. Die englischsprachige Wikipedia ist mit weit Ã¼ber drei Millionen Artikeln die grÃ¶Ã
deutschsprachigen Wikipedia mit Ã¼ber einer Million Artikeln.[1]

Die EintrÃ¤ge (â€žArtikelâ€œ u. a.) der Wikipedia werden von individuellen Autoren â€“ seltener von kollek
konzipiert, geschrieben und nach der VerÃ¶ffentlichung gemeinschaftlich korrigiert, erweitert und aktualisier

Das Ziel von Wikipedia ist es, eine frei lizenzierte und qualitativ hochstehende EnzyklopÃ¤die zu schaffen
Wikipedia nicht nur lesen, sondern auch als Autor mitwirken. Um Inhalte zu verÃ¤ndern, ist eine Anmeldun

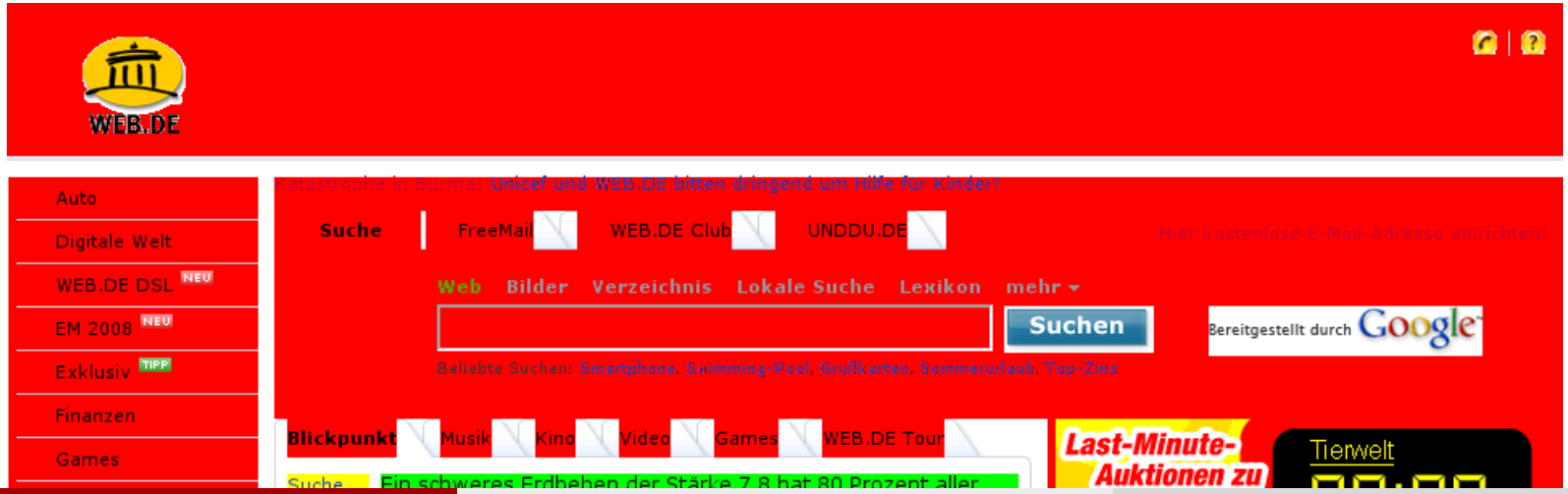
(Some) Things can go wrong – and if they can they will. . .

EURAC research

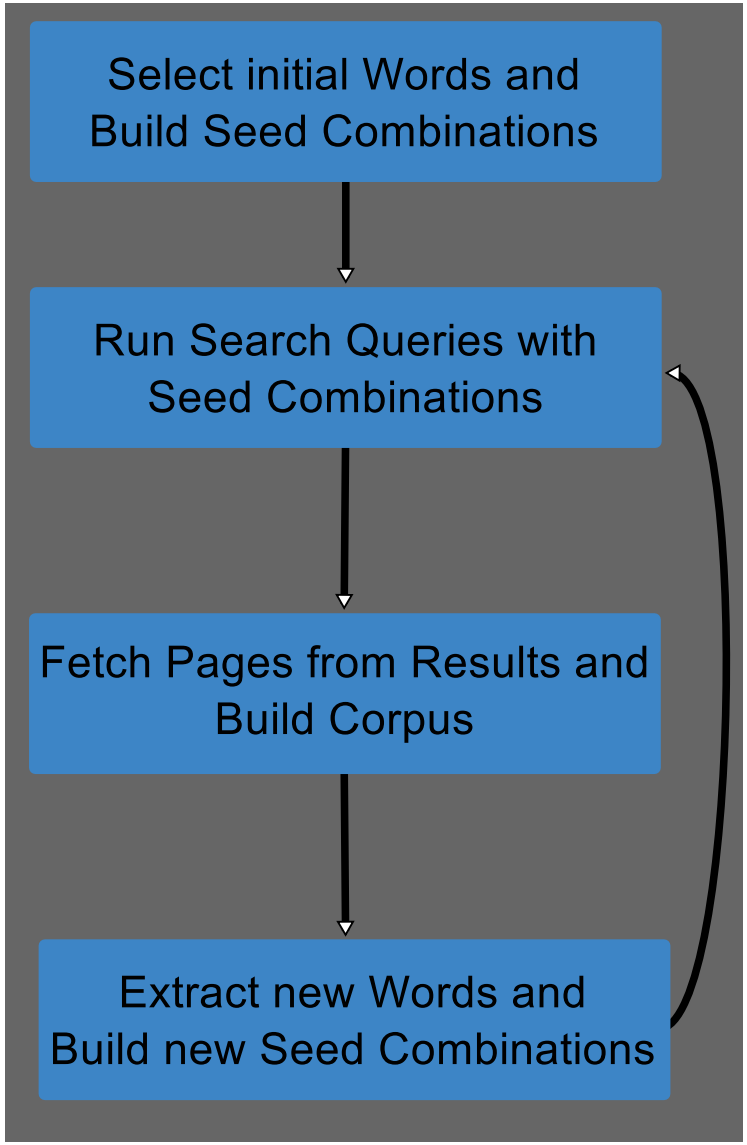
Documents might turn out to

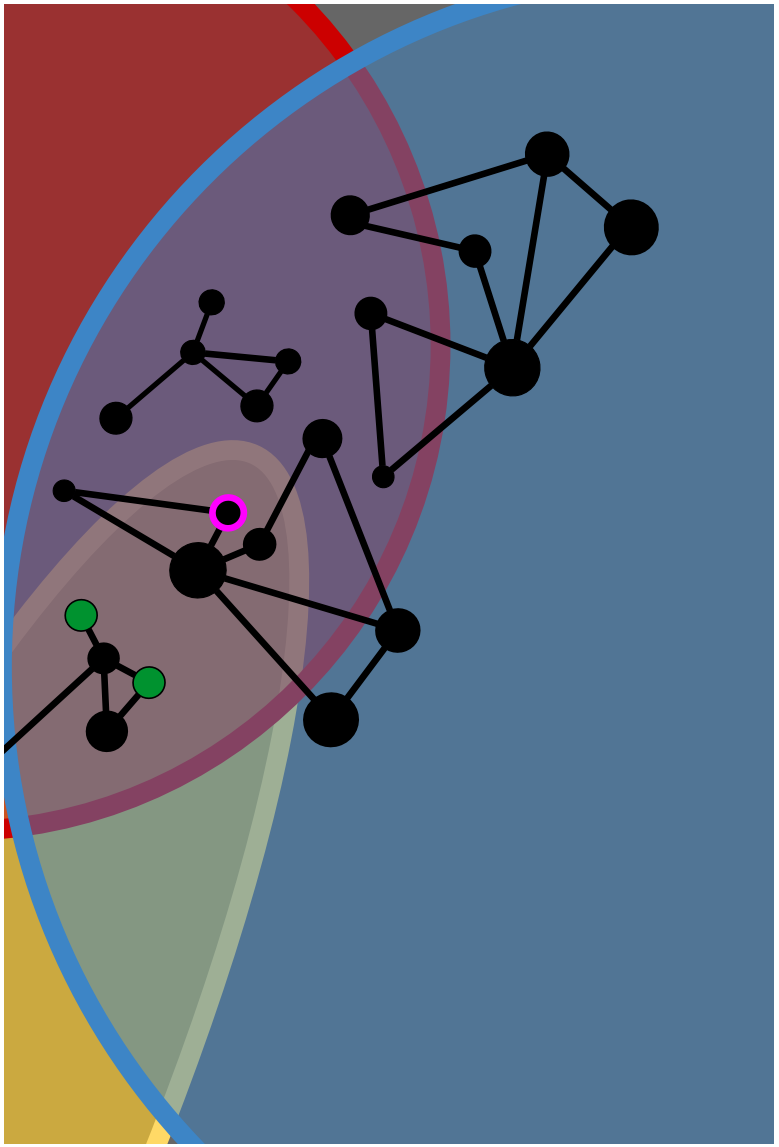
1. be in the wrong language (or of the wrong genre)
2. contain gibberish (characters, words, . . .)
3. contain *undesireable* content
   - parts of a page (e.g. boilerplate)
   - whole pages (e.g. duplicates or near-duplicates)
   - whole sites (e.g. bot-traps)

   . . .



The basic (BootCaT) Idea





How to select Words for the initial Seed Combinations

EURAC research

. . . or: it's good to have a corpus to build one

Use a small list (in the 5-to-15 range) of middle-frequency words from a general corpus.

Digression: For a *specialized corpus* words that are expected to be representative of this very domain can be used, e.g. names of rock bands.

How to run Search Queries

EURAC research

Application Programming Interface (API)

An API is a means for software to interact with other software.

Major search engines (e.g. Google, Yahoo!, Bing, Ask.com) provide APIs that let you specify (some of) the following features:

- the language of the result pages
- the country (or region) to which to restrict your search results, i.e. only results on web sites within this country are returned
- the Creative Commons license that the contents are licensed under

How to fetch Pages I

EURAC research

Web Crawler

A (web) crawler is a software agent (or bot) that browses the World Wide Web in a methodical, automated manner. It visits an initial list of seed URLs, identifies all the hyperlinks in the pages and adds them to a list of URLs still to visit.

Some characteristics of the web make crawling very difficult – crawlers take care of

- obeying politeness policies
  (visits, re-visits, parallelization, . . .)
- URL normalization
- naïve de-duplication (sometimes)

How to fetch Pages II

EURAC research

Duplicates

- Exact duplicates are exact copies — and easy to identify
- Near-duplicates are identical in terms of *content* but differ in a small portion of the document such as e.g., advertisement, counters, or date — and are more difficult to identify

De-duplication

1. Use a dimensionality reduction technique to map web page content to small sized fingerprints
2. Use *fingerprinting* that computes similar values for similar documents
3. Consider 'similar enough' fingerprints to represent similar documents

How to select Words for the new Seed Combinations

EURAC research

or: it's better to have your own corpus. . .

New seed words are extracted from the retrieved pages by comparing the frequency of occurrence of each word in this set with its frequency of occurrence in a reference corpus.

A BootCaT Example

EURAC research

- 7 seeds: black sabbath, led zeppelin, deep purple, motorhead, rainbow, judas priest, iron maiden
- 35 3-seed combinations:
  "led zeppelin" rainbow "black sabbath"
  "deep purple" motorhead rainbow
  "deep purple" "judas priest" motorhead
  . . .
- consider first 20 search results per query
- (ideally) 700 web pages

The big Picture (cont.)

EURAC research

Observation

However, after crawling content from the web the subsequent steps, namely, language identification, tokenising, lemmatising, part-of-speech tagging, indexing, etc. suffer from

> *'large and messy' training corpora [. . .] and interesting [. . .] regularities may easily be lost among the countless duplicates, index and directory pages, web spam, open or disguised advertising, and boilerplate.*

The Problem

Thorough pre-processing and cleaning of web corpora is crucial in order to obtain reliable frequency data.

What is a ‘clean’ Page?

Red: unwanted boilerplate; Yellow: Captions (titles, sub-titles, headings, etc.); Green: wanted running text.

Cleaning with a standard HTML Formatter

```
Blackmore's Night Latest News
Ritchie Blackmore's Bio
Blackmore's Night Band Bios
Blackmore's Night Tour Info
Blackmore's Night Merchandise
Blackmore's Night Photo Gallery
Blackmore's Night Audio Clips
...
Register for
Blackmores Night
Email Updates!
Just enter your
email address in
the box below and
click the 'Sign up' button!
...
RITCHIE BLACKMORE A MUSICAL HISTORY...
1967 - RITCHIE BLACKMORE - who has previously played with such bands
as the Outlaws, Screaming Lord Sutch, and Neil Christian & The
Crusaders - is invited by ex-Artwoods/The Flowerpot Men keybordist Jon
Lord (who was invited by The Searchers ex-drummer, Chris Curtis) to
form a new band. Other musician's would be auditioned from a Melody
Maker ad in Deeves Hall in Hertfordshire.
1968- In February, the group would form as Roundabout, consisting of
the three (with Chris Curtis on vocals) along with Dave Curtis on bass
and Bobby Woodman on drums. After only a month of uncompromising
rehearsals, BLACKMORE and LORD would be the only two remaining,
...
```

Cleaning with Finn's BTE Heuristic I

EURAC research

- Basic observation: Content-rich section of page tends to occur in low-HTML-density area
- Look for stretch that maximizes the quantity:
  N(TOKEN) - N(TAG)

Cleaning with Finn's BTE Heuristic II

EURAC research

```
<h2><a name="...">Background and motivation</a></h2>
<div class="level2">
<p>
<a href="link"><img src="source" alt="Motivational Poster" /></a>
</p>
<p>
Corpus-based distributional models (such as LSA or HAL)
have been claimed to capture interesting aspects of word meaning
...
</p>
```

Cleaning with Finn's BTE Heuristic II

EURAC research

TAG TAG TOKEN TOKEN TOKEN TAG TAG
TAG
TAG
TAG TAG TAG
TAG
TAG
TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
...
TAG

The big Picture's Solution: The Mission

EURAC research

Statement

The dimension of the cleaning task calls for an automated solution, the broadness of the problem for machine learning based approaches.

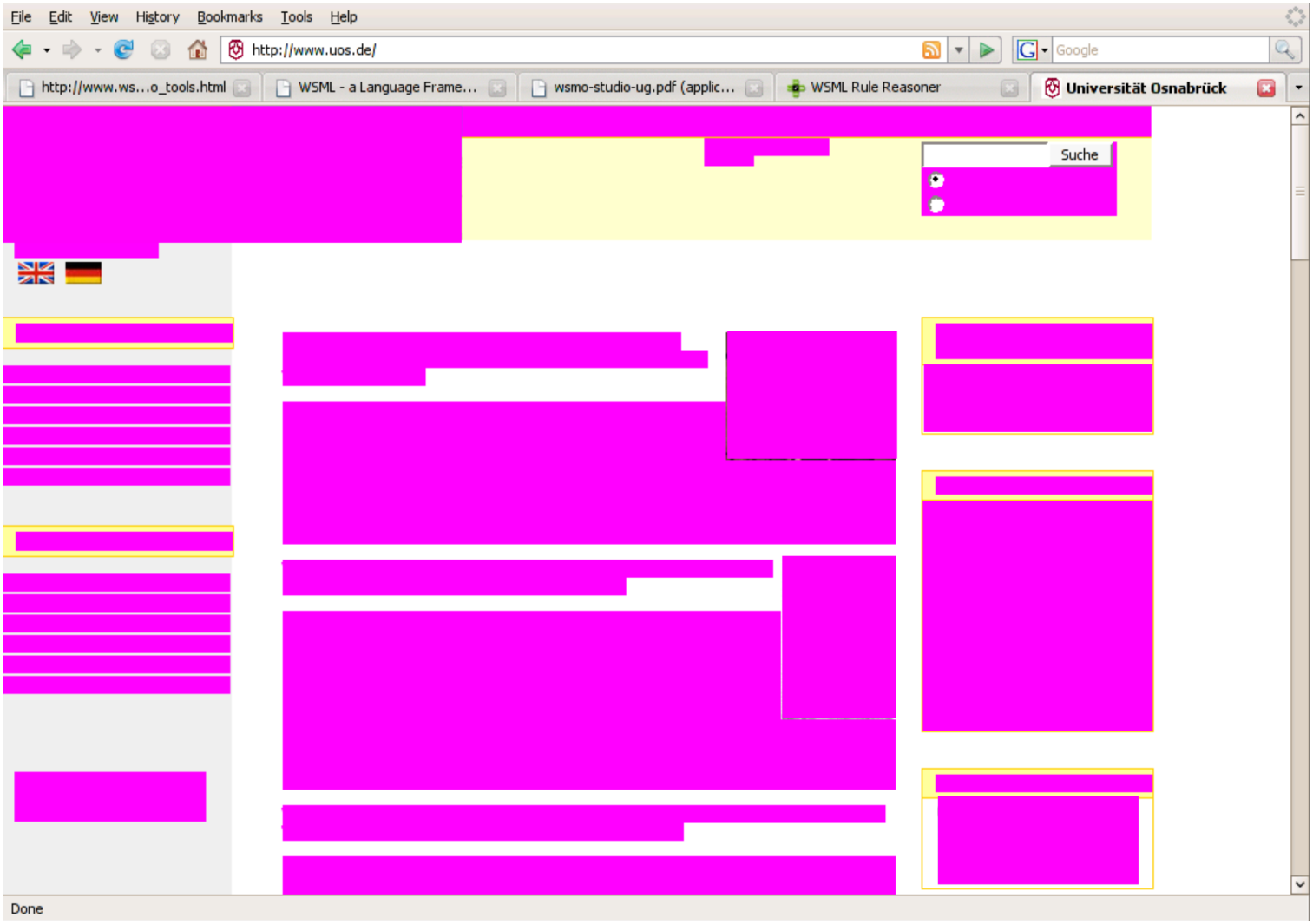
Observation

Part of the KrdWrd project deals with the development of appropriate methods, but they require hand-annotated pages for training.

A (smaller) new Problem

Develop a feasible way to tag web content.

. . . and another Approach to Cleaning



The KrdWrd Project

EURAC research

The KrdWrd Project includes a Firefox Add-on that facilitates the necessary tagging of web pages possible.

For users, we provide accurate page presentation and annotation utilities in a typical browsing environment, *while preserving* the original document and all the additional information contained therein.

. . . while still being at it:

preserve as much web content as possible.

The KrdWrd Project includes a Proxy Set-up: This storage fills up with the harvested web pages but also with all directly-linked material, which is included via absolute or relative links.

The Results I

EURAC research

...can be viewed here:
https://krdwrd.org/screencasts/cast01.html

...and here:
http://krdwrd.org

The Results II

EURAC research

The Gold Standard Corpus

- Length of documents was fixed between 500 and 6,000 words
- Final Data Set:
  - 219 web pages, consisting of more than 420,000 words and over 2.5 million characters, were
  - independently processed by 64 users who submitted
  - 1595 results (re-submits for a page counted only once), i.e.
  - an average of 7.28 submits/page.
- Average inter-coder agreement (Fleiss's multi-π) over all valid submissions is 0.85

The Results III

EURAC research

Table: *Weighted* 10-fold cross validated classification test results for different combinations of the textual (txt), DOM-property based (dom) and visual (viz) pipelines on the *Canola* (i.e. Gold Standard) data set.

Modules	Number of Features	Precision	Recall
txt	21	92%	**93%**
dom	13	89%	91%
viz	8	90%	**93%**
dom viz	21	90%	92%
txt viz	29	**94%**	**93%**
txt dom viz	42	93%	92%
BTE		80%	**99%**

Challenges 2.0

EURAC research

World Wide Web 2.0

The term "Web 2.0" was coined in January 1999 by Darcy DiNucci, a consultant on electronic information design (information architecture). In her article, "Fragmented Future", DiNucci writes:

“The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop. The Web will be understood not as screenfulls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will [...] appear on your computer screen, [...] on your TV set [...] your car dashboard [...] your cell phone [...] hand-held game machines [...] maybe even your microwave oven.”

References I

[@GO10] Google web search api [online].
2010.
Available from: GoogleWebSearchAPI.

[@KW08] Johannes M. Steger and Egon W. Stemle.
KrdWrd [online].
2008.
Available from: https://krdwrd.org.

[@YA10] YAHOO! developer network [online].
2010.
Available from: http://developer.yahoo.com/search/boss/boss_guide/overview.html.

[BB04] Marco Baroni and Silvia Bernardini.
BootCaT: Bootstrapping corpora and terms from the web.
In (ELRA) [EL04], pages 1313–1316.
Available from: http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf.

References II

[BBE08] Silvia Bernardini, Marco Baroni, and Stefan Evert.
A wacky introduction.
2008.
Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.126.5154.

[BCKS08] Marco Baroni, Francis Chantree, Kilgarriff, and Serge Sharoff.
CleanEval: A competition for cleaning web pages.
In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[BDD+07] Daniel Bauer, Judith Degen, Xiaoye Deng, Priska Herger, Jan Gasthaus, Eugenie Giesbrecht, Lina Jansen, Christin Kalina, Thorben Krüger, Robert Märtin, Martin Schmidt, Simon Scholler, Johannes Steger, Egon Stemle, and Stefan Evert.
FIASCO: Filtering the internet by automatic subtree classification, osnabrück.
In *Building and Exploring Web Corpora (WAC3 - 2007)*, 2007.

References III

[BS05] Marco Baroni and Serge Sharoff.
Creating specialized and general corpora using automated search engine queries.
Technical report, SSLMIT, University of Bologna; CTS, University of Leeds, 2005.
Available from: http://sslmit.unibo.it/~baroni/wac/serge_marco_wac_talk.slides.pdf.

[BU06] Marco Baroni and Motoko Ueyama.
Building general- and special-purpose corpora by web crawling.
In *Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application*, 2006.
Available from: http://explorer.csse.uwa.edu.au/reference/browse_paper.php?pid=233281973.

[EL04] European Language Resources Association (ELRA), editor.
Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, May 2004.
Available from: http://www.lrec-conf.org/lrec2004/.

References IV

[Eve08] Stefan Evert.
A lightweight and efficient tool for cleaning web pages.
In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[FKS01] A. Finn, N. Kushmerick, and B. Smyth.
Fact or fiction: Content classification for digital libraries.
In *Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries (Dublin)*, 2001.

[MJD07] Gurmeet Singh (Google Inc.) Manku, Arvind (Google Inc.) Jain, and Anish (Stanford University) Das Sarma.
Detecting near-duplicates for web crawling.
In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150, New York, New York, USA, 2007. ACM.
Available from: `http://portal.acm.org/citation.cfm?id=1242592`.

References V

[MS11] Brian Murphy and Egon W. Stemle.
PaddyWaC: A Minimally-Supervised Web-Corpus of Hiberno-English.
In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 22–29, Edinburgh, Scotland, UK, 2011. Association for Computational Linguistics.
Available from: `http://www.aclweb.org/anthology/W11-2603`.

[THG05] Jose Tummers, Kris Heylen, and Dirk Geeraerts.
Usage-based approaches in cognitive linguistics: A technical state of the art.
Corpus Linguistics and Linguistic Theory, 1(2):225–261, 11 2005.