

# Terminology Work — Breaking the Barriers

# **Automatic Term Extraction**

An Stuyven – Skrivanek Group

Vancouver, October 29, 2014



# Experience with Terminology



Skrivanek has been doing for 20 years

⇒ terminology work with small and large clients

⇒ covering the whole terminology workflow:



- Terminology Process Consultancy
- Clearing the objectives and conditions (existing terminology, type and size of the content, quality requirements, tools, integration, cost, ...)
- Input and validation of existing glossaries
- Terminology extraction and translation
- Online terminology sharing (tools) with whole corporation
- Terminology maintenance and ongoing processing





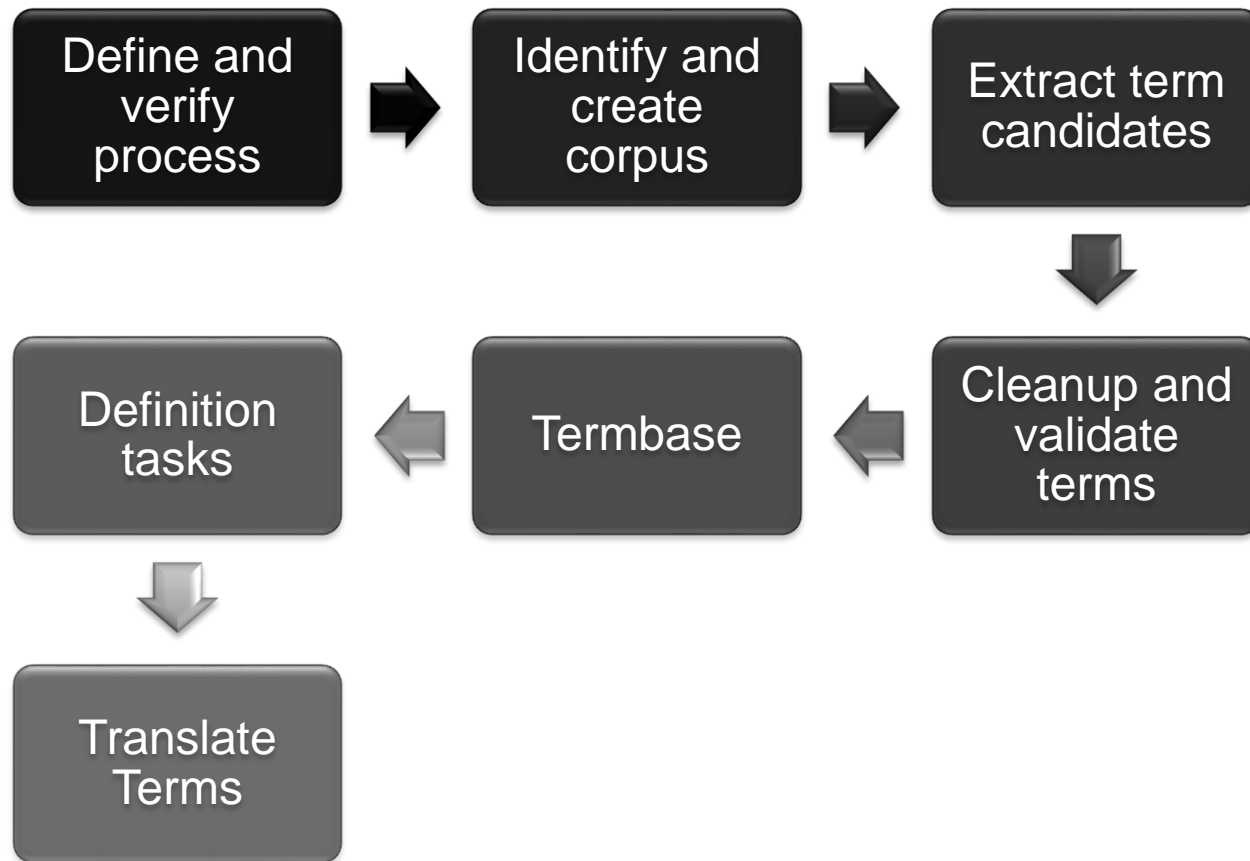
# Typical Barriers



- Input
- Varying candidate quality
- Output
- Compatibility
- User-friendliness of sharing method



# Term Creation and Validation





# Term extraction



- Manual extraction



- Automatic extraction: tools

e.g.: TerMine,

+ Most TMS tools have this function (SDL Trados, MemoQ,...)



- Main problem:

high percentage of „noise and silence“





# Term Extraction



## Statistical Approach

### Most common

Examples: TerMine, Fivefilters Term Extraction, SDL Multiterm Extract, ...



### Based on frequency

- Language independent
- Issues
  - The frequency threshold must be specified
  - Frequency does not necessarily mean importance
  - Much „noise“ and „silence“ – extensive manual postprocessing



## Linguistic Approach

### Based on rules and dictionaries

- Not available for all languages
- Issues
  - Loans
  - Synonyms, variants, abbreviations
  - Ellipses
  - Improper usage

# Term extraction with qTerm (MemoQ)

Extract candidates

**Session name** SMB\_EN-DE\_TB extraction\_1

Sources

Translation documents  
 Translation memories  
 LiveDocs corpus documents

Every document  
 All memories in project  
 All documents shown

Selected documents  
 Primary TM  
 Selected TMs

Options

**General**

Maximum length (words)   
Minimum frequency   
Expression delimiters   
Length factor   
 Ignore words with numbers

**Single-word terms**

Minimum length (characters)   
Minimum frequency

**Term base lookup**

Look up candidates  
 All term bases in project  
 Term base with the highest rank only

Stop words

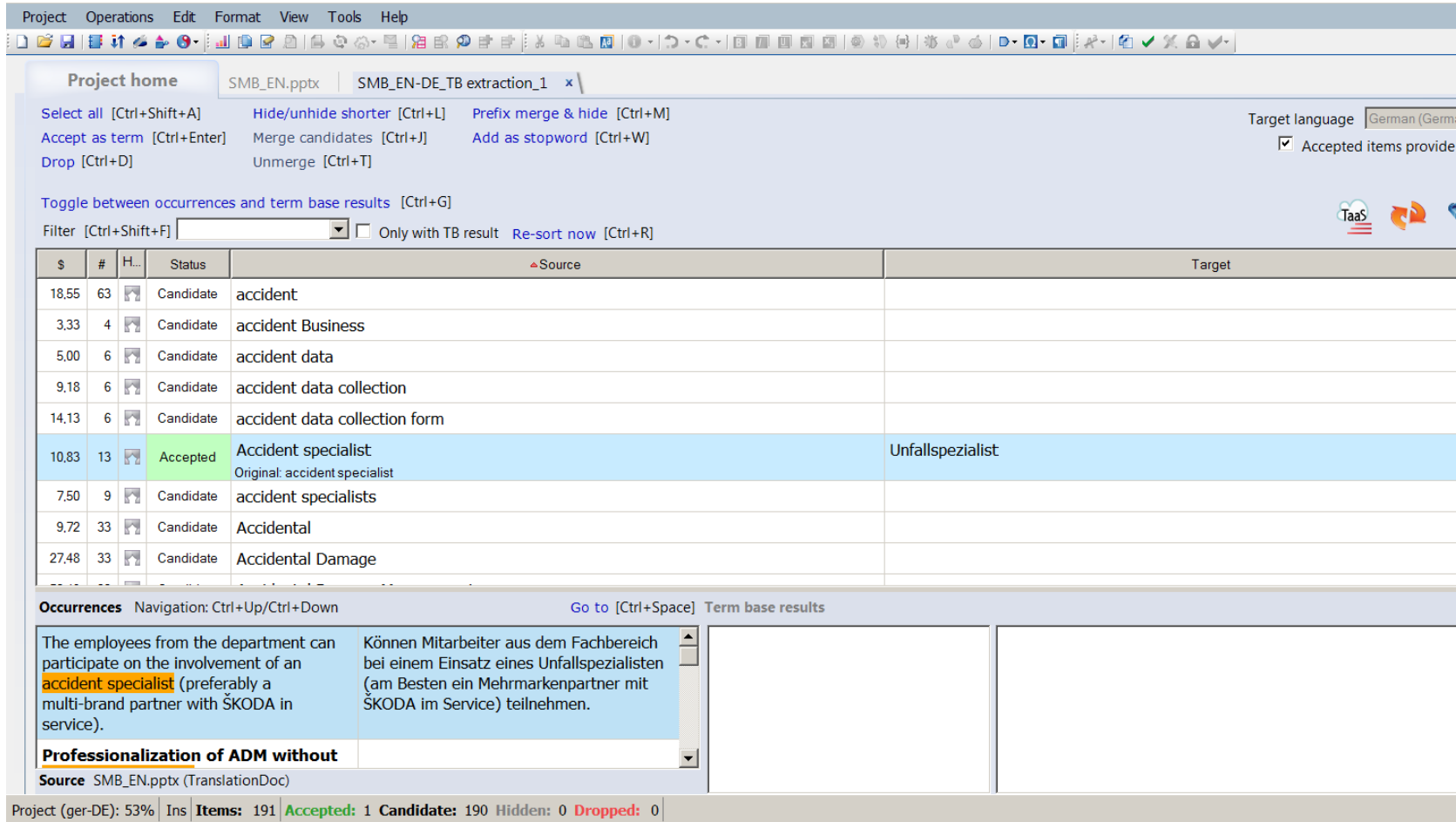
**Stop word list**  Save as...

Word	Blocks as first	Blocks inside	Blocks as last
another	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
any	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
are	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
about	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
above	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Word  Add Delete selected

OK Cancel Help

# Term extraction with qTerm -Candidates



The screenshot displays the qTerm software interface. The main window shows a list of term candidates with columns for score, frequency, status, source, and target. The candidate 'Accident specialist' is highlighted as 'Accepted' with a target of 'Unfallsspezialist'. Below the list, the 'Occurrences' section shows the source text for the selected term, with the term highlighted in yellow. The status bar at the bottom indicates the project progress: 53% complete, 191 items, 1 accepted, 190 candidates, 0 hidden, and 0 dropped.

\$	#	H...	Status	Source	Target
18,55	63		Candidate	accident	
3,33	4		Candidate	accident Business	
5,00	6		Candidate	accident data	
9,18	6		Candidate	accident data collection	
14,13	6		Candidate	accident data collection form	
10,83	13		Accepted	Accident specialist Original: accident specialist	Unfallsspezialist
7,50	9		Candidate	accident specialists	
9,72	33		Candidate	Accidental	
27,48	33		Candidate	Accidental Damage	

**Occurrences** Navigation: Ctrl+Up/Ctrl+Down      Go to [Ctrl+Space] Term base results

The employees from the department can participate on the involvement of an **accident specialist** (preferably a multi-brand partner with ŠKODA in service).      Können Mitarbeiter aus dem Fachbereich bei einem Einsatz eines Unfallsspezialisten (am Besten ein Mehrmarkenpartner mit ŠKODA im Service) teilnehmen.

**Professionalization of ADM without**

Source SMB\_EN.pptx (TranslationDoc)

Project (ger-DE): 53% Ins Items: 191 Accepted: 1 Candidate: 190 Hidden: 0 Dropped: 0



# Term extraction with SDL Multiterm (Trados)

The screenshot displays the SDL Multiterm (Trados) software interface. The main window shows a list of terms extracted from a document, organized into two columns: English (United Kingdom) and German (Germany). The terms are sorted by score, with the highest score being 99. The term 'Accident Specialist' is highlighted in the English column, and its corresponding German translation 'Unfallspezialisten' is checked in the German column. Below the main list, the 'Term properties' window is open, showing the details for the selected term. The 'Domain' is set to '<None>', and the 'Synonyms' field is empty. The 'Word forms' field contains 'specialists,Accident Specialist,accident specialist'. The 'Filename' is 'SMB\_EN\_DE.tmx;'. The 'Concordance' window is also open, showing 0 sentences.

Score	Domain	English (United Kingdom)	German (Germany)
99	<None>	<input type="checkbox"/> VW	<input type="checkbox"/> VW
99	<None>	<input type="checkbox"/> Service Plans	<input type="checkbox"/> Service Programme
			<input type="checkbox"/> Unfallschadenmanagement
99	<None>	<input type="checkbox"/> dealership	<input type="checkbox"/> Autohaus
			<input type="checkbox"/> Betrieben
99	<None>	<input type="checkbox"/> Clever Repair	<input type="checkbox"/> Clever Repair
99	<None>	<input checked="" type="checkbox"/> Accident Specialist	<input checked="" type="checkbox"/> Unfallspezialisten
99	<None>	<input type="checkbox"/> accident data collection form	<input type="checkbox"/> per Fax
			<input type="checkbox"/> Unfallerfassungsbogen ausfüllen
99	<None>	<input type="checkbox"/> accidental damage management	<input type="checkbox"/> Unfallschadenmanagement
			<input type="checkbox"/> Service Programme
89	<None>	<input type="checkbox"/> ŠKODA	<input type="checkbox"/> ŠKODA
88	<None>	<input type="checkbox"/> Involvement of an Accident	<input type="checkbox"/> Einsatz eines Unfallspezialisten
86	<None>	<input type="checkbox"/> ADM	<input type="checkbox"/> USM
84	<None>	<input type="checkbox"/> Training Offer	<input type="checkbox"/> Schulungsangebot

**Term properties**

Accident Specialist

Unfallspezialisten

<Click here to enter a new translation>

Domain: <None>

Synonyms: <Click here to enter a new synonym>

Acronym:

Word forms: specialists,Accident Specialist,accident specialist

Filename: SMB\_EN\_DE.tmx;

Definition:

Note:

Context sentences:  Add new sentence  Remove sentence  Generate sentences

Created: 20.10.14 8:29:51 Modified: 20.10.14 8:30:14



# Sketchengine



<http://www.sketchengine.co.uk/>



The Sketch Engine is for anyone wanting to research how words behave. It is a Corpus Query System



Concordance

Word sketches



**goal** (*noun*) ukWaC freq = 168345 (107.5 per million)

<u>object of</u>	<u>58924</u>	<u>3.2</u>	<u>subject of</u>	<u>25451</u>	<u>2.4</u>	<u>modifier</u>	<u>67879</u>	<u>1.6</u>	<u>modifies</u>	<u>11026</u>	<u>0.3</u>
score	<a href="#">8390</a>	11.28	score	<a href="#">903</a>	8.59	ultimate	<a href="#">1911</a>	9.27	scorer	<a href="#">389</a>	9.39
achieve	<a href="#">9422</a>	9.9	disallow	<a href="#">223</a>	8.04	long-term	<a href="#">875</a>	7.66	kick	<a href="#">634</a>	8.86
concede	<a href="#">1421</a>	9.39	concede	<a href="#">204</a>	7.53	league	<a href="#">638</a>	7.38	tally	<a href="#">129</a>	7.9
accomplish	<a href="#">585</a>	7.97	gape	<a href="#">76</a>	6.5	winning	<a href="#">401</a>	7.33	keeper	<a href="#">204</a>	7.31
reach	<a href="#">1924</a>	7.66	come	<a href="#">1316</a>	5.44	primary	<a href="#">993</a>	7.24	scramble	<a href="#">50</a>	6.75
net	<a href="#">337</a>	7.42	kick	<a href="#">76</a>	5.44	second	<a href="#">2000</a>	7.19	drought	<a href="#">78</a>	6.65
pursue	<a href="#">648</a>	7.41	rule	<a href="#">61</a>	5.24	common	<a href="#">1529</a>	7.17	difference	<a href="#">676</a>	6.28
attain	<a href="#">400</a>	7.35	orientate	<a href="#">34</a>	5.06	strategic	<a href="#">645</a>	7.1	cushion	<a href="#">53</a>	6.26
grab	<a href="#">406</a>	7.34	arrive	<a href="#">90</a>	4.43	realistic	<a href="#">422</a>	7.05	lead	<a href="#">267</a>	6.24
set	<a href="#">2413</a>	7.01	cap	<a href="#">20</a>	4.38	achievable	<a href="#">290</a>	6.97	setting	<a href="#">405</a>	6.14
pull	<a href="#">501</a>	6.88	beat	<a href="#">53</a>	4.31	stated	<a href="#">259</a>	6.8	kicker	<a href="#">25</a>	6.04
disallow	<a href="#">190</a>	6.67	direct	<a href="#">53</a>	4.22	score	<a href="#">611</a>	6.75	post	<a href="#">482</a>	5.91



# Term Finding with Sketchengine



## **Linguistic approach**

Tokenising, lemmatising, POS-tagging

+



## **Statistical approach**

Frequency in domain corpus vs reference corpus



- Around 70 languages
- Company is specialized in building these reference corpora



# Term Finding with Sketchengine



Term candidates for a domain, in a language, can be found by:



- Taking a corpus for the domain and a reference corpus for the language
- Identifying the grammatical shape of a term in the language
- Tokenising, lemmatising and POS-tagging both corpora
- Identifying and counting the items in each corpus which match the grammatical shape
- For each item in the domain corpus, comparing its frequency with its frequency in the reference corpus



-> Items with highest domain:reference ratio  
are the top term candidates

## Corpora















- [+ Create corpus](#)
- [+ WebBootCaT](#)
- [+ Upload TMX](#)

- [Parallel corpora](#)
- [Compare corpora](#)
- [Configuration templates](#)
- [Sketch grammars](#)
- [Subcorpus definitions](#)
- [User groups](#)

Support






- [Help index](#)
- [Report an error](#)
- [Request a feature](#)

## Corpora

Language	Corpus name	Tokens	Words	
Chinese, Simplified	<a href="#">zhTenTen11</a>	2,106,661,021	1,729,867,455	 
English	<a href="#">British National Corpus</a>	112,181,015	96,048,950	 
English	<a href="#">enTenTen [2012]</a>	12,968,375,937	11,191,860,036	 
French	<a href="#">frTenTen [2012]</a>	12,369,868,562	10,666,617,369	 
German	<a href="#">deTenTen [2010]</a>	2,844,839,761	2,338,036,362	 
Japanese	<a href="#">jpTenTen11 [SUW]</a>	10,321,875,664	8,432,256,386	 
Spanish	<a href="#">esTenTen11 (European, Freeling)</a>	2,341,159,406	1,991,879,282	 

[Featured corpora](#) | [All corpora](#) | [Parallel corpora](#)

## My corpora

Language	Corpus name	Configuration template	Tokens	
English	<a href="#">EN-DE_TB Extraction</a>	TreeTagger for English	0	    

[+ Create corpus](#) | [+ WebBootCaT](#)



**Corpora**

- [+ Create corpus](#)
- [+ WebBootCaT](#)
- [+ Upload TMX](#)

**Parallel corpora**

[Compare corpora](#)

**Configuration templates**

[Sketch grammars](#)

[Subcorpus definitions](#)

[User groups](#)

Corpus

- [Corpus page](#)
- [+ Add new file](#)
- [+ Add web data \(BootCaT\)](#)
- [⌚ Compile corpus](#)
- [🔍 Search corpus](#)
- [📁 Extract keywords & terms](#)
- [🔧 Configure corpus](#)
- [🔧 Change sketch grammar](#)
- [🔧 Set subcorpus definitions](#)
- [🚫 Expert mode](#)
- [📂 Download corpus](#)
- [🔑 Access privileges](#)
- [🔍 View logs](#)

Support

[Help index](#)

[Report an error](#)

## EN-DE\_TB Extraction: Extracted keywords and terms

[Change extraction options](#)

Keywords	Score	F	RefF	Terms	Score	F	RefF
<input type="checkbox"/> škoda	9,628.64	<a href="#">27</a>	<a href="#">860</a>	<input type="checkbox"/> accidental damage	3,006.55	<a href="#">10</a>	<a href="#">3,443</a>
<input type="checkbox"/> adm	3,359.92	<a href="#">13</a>	<a href="#">6,114</a>	<input type="checkbox"/> accident specialist	2,644.09	<a href="#">7</a>	<a href="#">94</a>
<input type="checkbox"/> autorechtaktuell	1,521.91	<a href="#">4</a>	<a href="#">0</a>	<input type="checkbox"/> collection form	1,874.03	<a href="#">5</a>	<a href="#">206</a>
<input type="checkbox"/> ohne	1,322.00	<a href="#">5</a>	<a href="#">5,691</a>	<input type="checkbox"/> data collection form	1,518.87	<a href="#">4</a>	<a href="#">32</a>
<input type="checkbox"/> accidental	1,276.94	<a href="#">23</a>	<a href="#">75,857</a>	<input type="checkbox"/> authorization agreement	1,512.83	<a href="#">4</a>	<a href="#">89</a>
<input type="checkbox"/> usm	1,043.86	<a href="#">4</a>	<a href="#">5,939</a>	<input type="checkbox"/> damage management	1,132.62	<a href="#">3</a>	<a href="#">115</a>
<input type="checkbox"/> aggrieved	887.59	<a href="#">4</a>	<a href="#">9,268</a>	<input type="checkbox"/> claim settlement	1,086.28	<a href="#">3</a>	<a href="#">672</a>
<input type="checkbox"/> constructio	750.69	<a href="#">2</a>	<a href="#">186</a>	<input type="checkbox"/> work accident	1,059.08	<a href="#">3</a>	<a href="#">1,022</a>
<input type="checkbox"/> cooperations	716.64	<a href="#">2</a>	<a href="#">811</a>	<input type="checkbox"/> external service	1,058.09	<a href="#">3</a>	<a href="#">1,030</a>
<input type="checkbox"/> bads	710.81	<a href="#">2</a>	<a href="#">924</a>	<input type="checkbox"/> data collection	943.53	<a href="#">4</a>	<a href="#">7,961</a>
<input type="checkbox"/> cession	674.26	<a href="#">2</a>	<a href="#">1,677</a>	<input type="checkbox"/> quota accident	761.46	<a href="#">2</a>	<a href="#">1</a>
<input type="checkbox"/> wolfsburg	672.61	<a href="#">2</a>	<a href="#">1,713</a>	<input type="checkbox"/> fax authorization agreement	761.46	<a href="#">2</a>	<a href="#">0</a>
<input type="checkbox"/> dealership	670.26	<a href="#">11</a>	<a href="#">67,975</a>	<input type="checkbox"/> collection form assignment	761.46	<a href="#">2</a>	<a href="#">0</a>
<input type="checkbox"/> seeger	572.91	<a href="#">2</a>	<a href="#">4,268</a>	<input type="checkbox"/> data collection form assignment	761.46	<a href="#">2</a>	<a href="#">0</a>
<input type="checkbox"/> plausibility	555.63	<a href="#">2</a>	<a href="#">4,804</a>	<input type="checkbox"/> form assignment	761.46	<a href="#">2</a>	<a href="#">6</a>
<input type="checkbox"/> trainings	526.15	<a href="#">5</a>	<a href="#">33,915</a>	<input type="checkbox"/> procedure audit	761.46	<a href="#">2</a>	<a href="#">2</a>
<input type="checkbox"/> jens	488.31	<a href="#">2</a>	<a href="#">7,254</a>	<input type="checkbox"/> residual value plausibility check	761.46	<a href="#">2</a>	<a href="#">0</a>
<input type="checkbox"/> accident	482.83	<a href="#">59</a>	<a href="#">589,597</a>	<input type="checkbox"/> value plausibility check	761.46	<a href="#">2</a>	<a href="#">0</a>
<input type="checkbox"/> incl	460.66	<a href="#">2</a>	<a href="#">8,468</a>	<input type="checkbox"/> accident data collection	761.46	<a href="#">2</a>	<a href="#">0</a>



# SketchEngine

## Building a domain corpus



WebBootCaT



If no domain corpus is available, it can be created:

- Send „seed terms“ to a commercial search engine
- Gather the indexed pages
- Cleaning, deplicating and indexing as a corpus
- Corpus can be used for translators to find concordance examples





Corpora

- + Create corpus
- + WebBootCaT
- + Upload TMX

Parallel corpora  
Compare corpora

- Configuration templates
- Sketch grammars
- Subcorpus definitions
- User groups

Corpus

- Corpus page
- + Add new file
- + Add web data (BootCaT)
- ⌚ Compile corpus
- 🔍 Search corpus
- 📁 Extract keywords & terms
- 🔧 Configure corpus
- 🔧 Change sketch grammar
- 🔧 Set subcorpus definitions
- 🚫 Expert mode
- 📄 Download corpus
- 🔑 Access privileges
- 🔍 View logs

# EN-DE\_TB Extraction: WebBootCaT

Name of the collection

Unique identifier of the data collection. May only contain letters, numbers, underscores.

Input type

Seed words

URLs

Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

Seed words

Random tuples will be selected from the seed words to query a search engine. Input 3 to 20 words or multiword expressions. Use space as separator. Enclose multiword expressions into quotes (").

Compile corpus when finished

Automatically compile corpus when WebBootCaT processing is finished.

[Show advanced options](#)

Corpora

- + Create corpus
- + WebBootCaT
- + Upload TMX

Parallel corpora  
Compare corpora  
Configuration templates  
Sketch grammars  
Subcorpus definitions  
User groups

Corpus

- Corpus page
- + Add new file
- + Add web data (BootCaT)
- ⌚ Compile corpus
- 🔍 Search corpus
- 📁 Extract keywords & terms

# EN-DE\_TB Extraction: WebBootCaT

Select URLs to download

Query: accident accidental cooperations

- <http://www.erv.ch/fr-CH/clients-privés/accident-et-maladie/assurance-maladie-monde-entier/international-healthcare/international-healthcare-en>
- <http://www.cs.umd.edu/~nau/papers/au2006accident.pdf>
- [http://www.jaea.go.jp/04/turuga/internationalworkshop/presentationPDF/201206130940\\_Nicolas%20Devictor\\_France.pdf](http://www.jaea.go.jp/04/turuga/internationalworkshop/presentationPDF/201206130940_Nicolas%20Devictor_France.pdf)
- <http://www.cs.utexas.edu/~chiu/papers/Au07accidental.pdf>
- [http://www.irsn.fr/EN/newsroom/News/Pages/20060831\\_gamma\\_radiography\\_accident\\_in\\_Africa\\_IRSN\\_provides\\_assistance](http://www.irsn.fr/EN/newsroom/News/Pages/20060831_gamma_radiography_accident_in_Africa_IRSN_provides_assistance)
- [http://www.mak.at/en/program/event?article\\_id=1350932813534](http://www.mak.at/en/program/event?article_id=1350932813534)
- [http://www.irsn.fr/EN/newsroom/News/Pages/20121200\\_Dest](http://www.irsn.fr/EN/newsroom/News/Pages/20121200_Dest)






# Rule-Based MT for Term Extraction

- 
- Rule based Machine Translation Systems can be used for Term Extraction\*

- 
- No direct function

Use the reverse: „unknown words“ list

- 
- Result depends on specialization of your topic and on MT feed



# Term Extraction



- Tool choice according to requirements (bilingual extraction,



- play with settings
  - input (corpus, TM, ...)
  - frequency
  - length of terms
  - stop word lists
  - ...





**Thank you for your attention!**

An Stuyven  
Skrivanek Group

[an.stuyven@skrivanek.com](mailto:an.stuyven@skrivanek.com)