Term Extraction

zerfass@zaac.de

memoq fest
29 - 31 May 2019 | Budapest, Hungary

Agenda

The Basics
- Manual extraction
- Term extraction feature

The Process
Working through the
list of candidates
Numbers and own experience

The Settings
Extraction settings

The Wish list

memoq fest

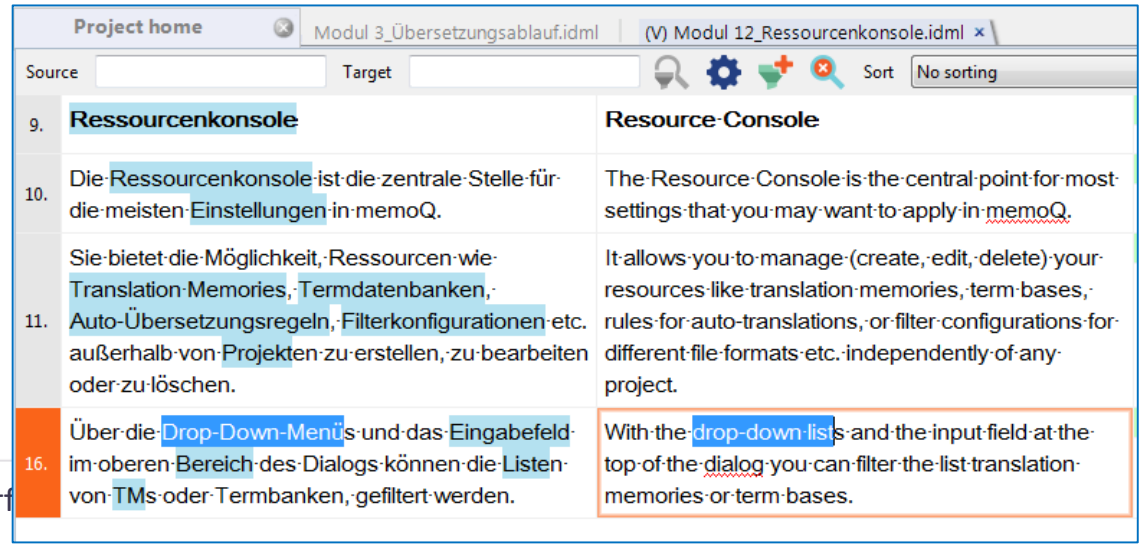
The Basics

Ways to extract terms

zerfass@zaac.de

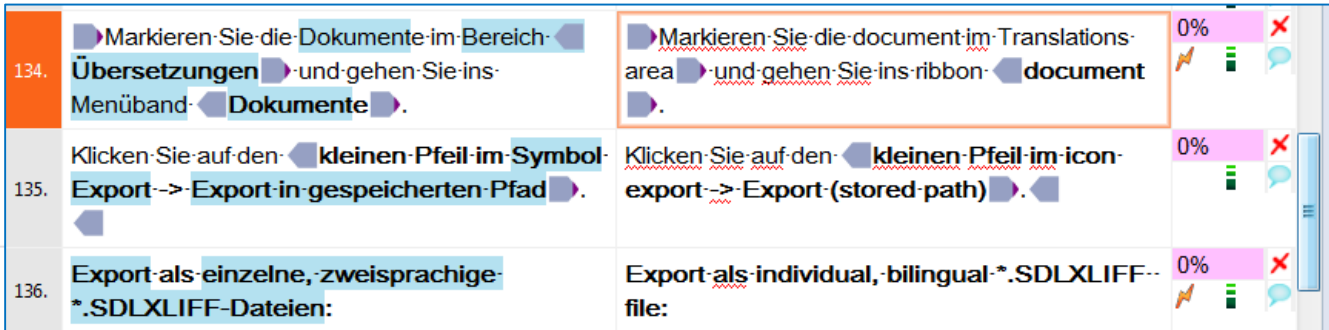
The manual process

- Set up a project with a term base
- Open the document (or documents as a view) in the editor.
- Click through all segments and see how many terms are already found in the term base (marked in blue)
- Add terms manually with CTRL-Q (term pairs if you go through an already translated document)



The manual process

- Angelika's experience:
  - If your documents contain up to 20,000 words in total the process to collect terms manually takes about as much time as going through a candidates' list from the extraction.
- To minimize the number of segments to go through in the editor:
  - Run a pre-translation with the fragment assembling using a term base and non-translatable list only (without TM). These segments already contain known terms and will probably not contain many more term candidates.

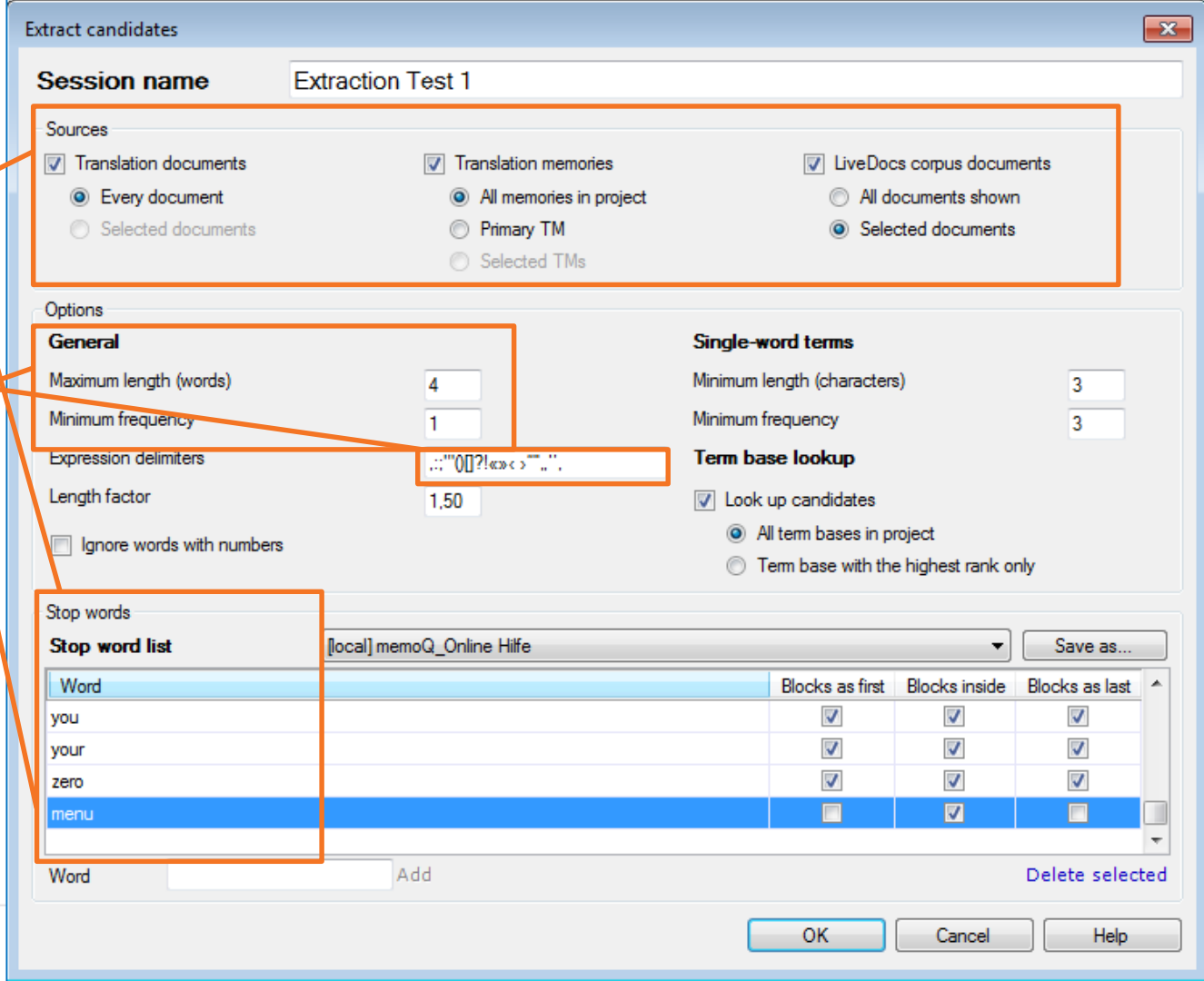


Term Extraction Module

- Statistical extraction from source language
- From
  - Translation documents
  - TMs
  - LiveDocs corpora (alignments, translation documents saved to LiveDocs and monolingual documents)
- Exports to
  - term bases
  - Excel
  - TaaS (term plug-in)

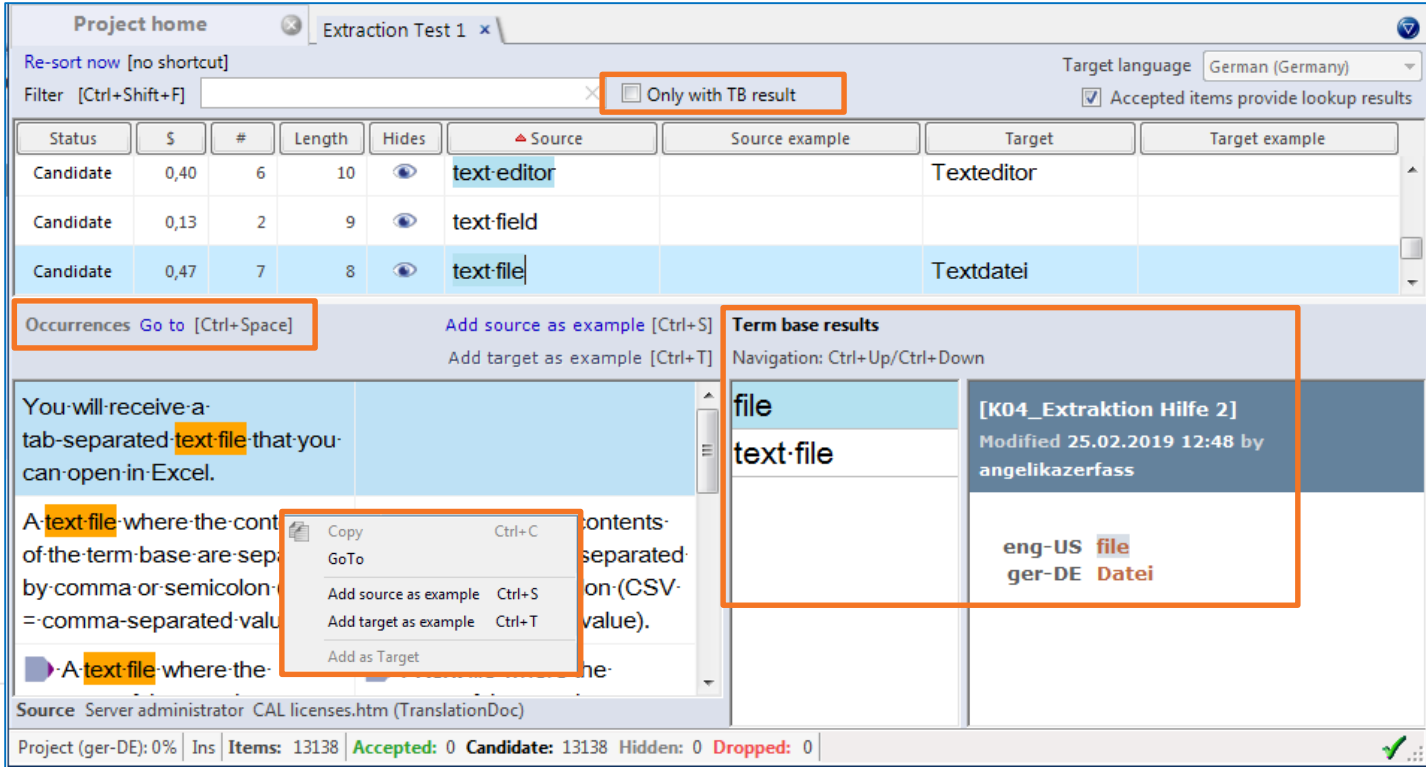
The Settings

- Check the text for the characters that come up between terms.
- Block as…
  - Finds "menu" as part of a term only, if the word "menu" appears at the beginning or end of the term *menu item, edit menu*



The List of Candidates

- Existing terms marked in blue / show only candidates with known terms
- Term base entries visible
- Go to feature to jump to source file
- Right-click to add segments as examples



The List of Candidates

- Accept or drop a term
- Edit the terms
- Sort the list by
  - the number of characters within the term
  - the frequency of terms/term combinations
  - alphabetically
- Export accepted terms to term base or Excel
- Send dropped single-word terms to the stop word list

memoq fest

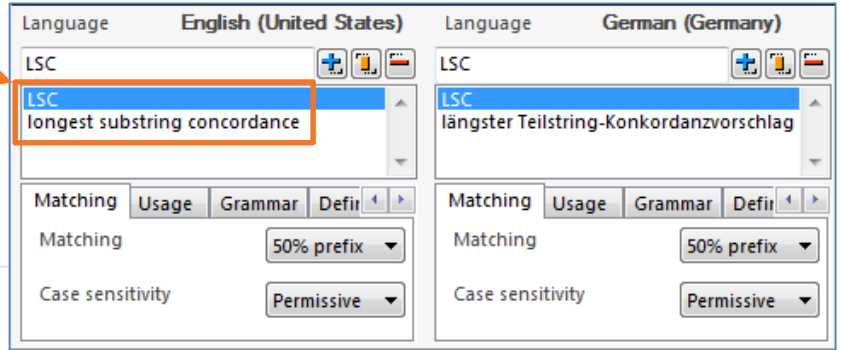
The Work in the Extraction List

- Run several extractions with different settings (settings cannot be changed once the extraction has run on the files).
- Accept AND drop terms as you go. If you don't mark terms as dropped and then re-sort the list, you might have to look at the same entries twice.

The Work in the Extraction List

- To add a synonym, use a semicolon and a space to separate the terms.

Status	$	#	Length	Hides	Source
Accepted	1,28	41	9		licensing
Accepted	0,18	2	15		licensing·method
Accepted	0,32	2	19		LSC;·longest·substring·concordance



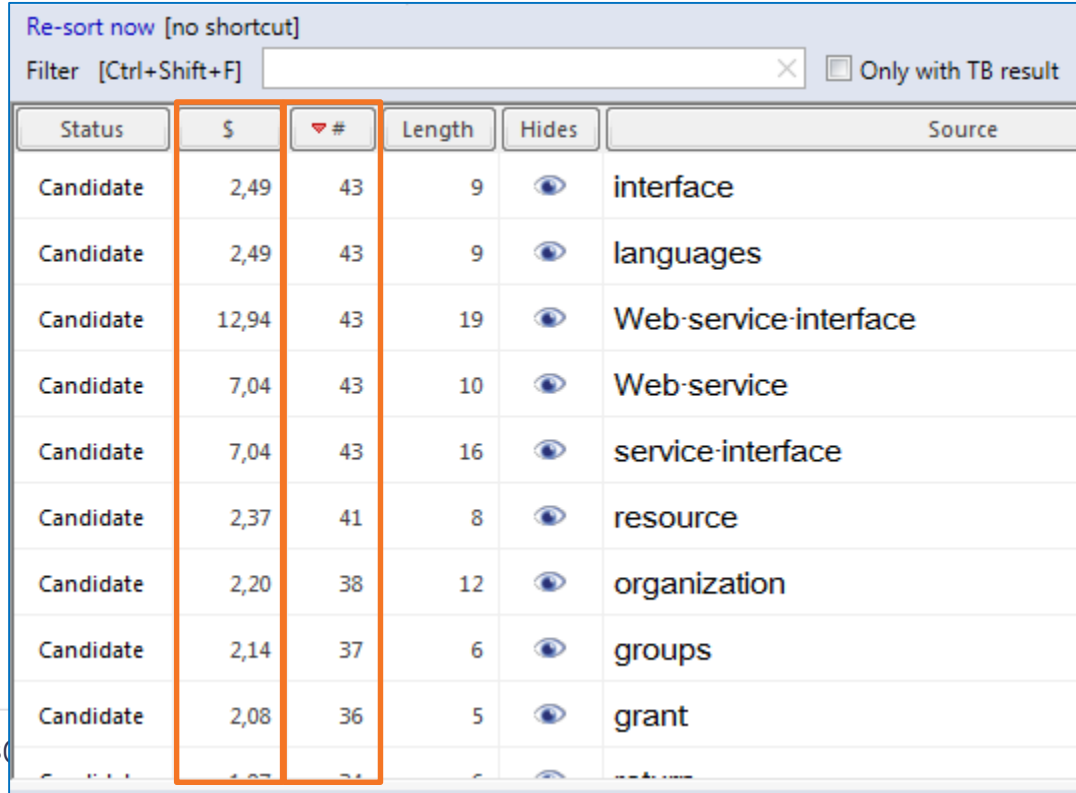
The Work in the Extraction List

- Sort by length
  - The longer the term, the higher the probability that it is a good term candidate
- Sort alphabetically
  - This also helps to find slightly different spellings of a term or highlight general inconsistencies, like the use of hyphens

Status	$	#	Length	Hides	Source
Candidate	0,08	1	15		Leasingabkommen
Candidate	0,08	1	16		Leasing-Laufzeit
Candidate	0,08	1	22		Leasingverpflichtungen
Candidate	0,25	3	15		Leasingverträge
Candidate	0,08	1	16		Leasing-Verträge
Candidate	0,08	1	16		Leasingzahlungen

The Work in the Extraction List

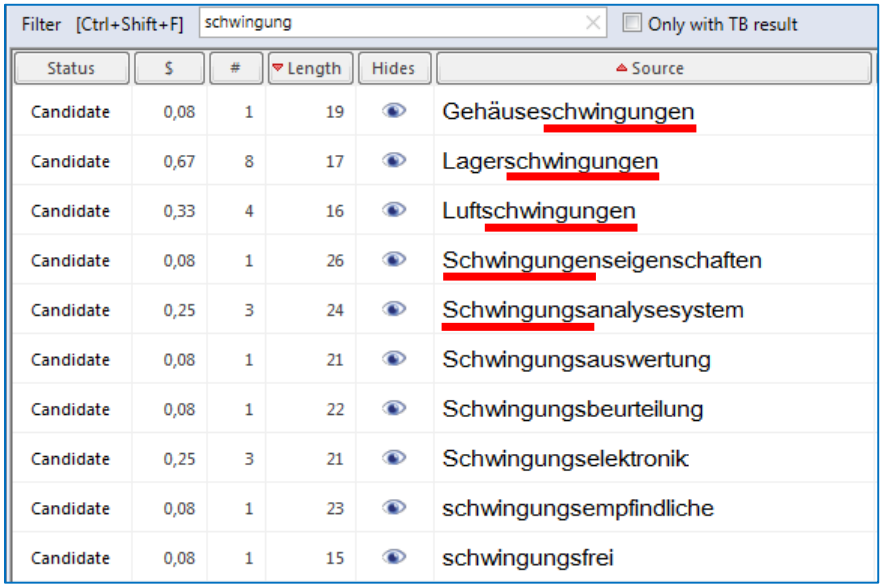
- Sort by frequency
  - If a full extraction is not possible (restraints on time/budget), concentrate on the most frequent terms.
- Sort by memoQ indicator ($)
  - The higher the number, the more memoQ thinks that this is a good candidate.



Status	$	#	Length	Hides	Source
Candidate	2,49	43	9		interface
Candidate	2,49	43	9		languages
Candidate	12,94	43	19		Web-service-interface
Candidate	7,04	43	10		Web-service
Candidate	7,04	43	16		service-interface
Candidate	2,37	41	8		resource
Candidate	2,20	38	12		organization
Candidate	2,14	37	6		groups
Candidate	2,08	36	5		grant

The Work in the Extraction List

- Filter for a specific part
  - Find all terms that contain this part
  - Drop all terms you don't want (Prevents the user from trying to confirm the same term several times)



Status	$	#	Length	Hides	Source
Candidate	0,08	1	19		Gehäuseschwingungen
Candidate	0,67	8	17		Lagerschwingungen
Candidate	0,33	4	16		Luftschwingungen
Candidate	0,08	1	26		Schwingungenseigenschaften
Candidate	0,25	3	24		Schwingungsanalysesystem
Candidate	0,08	1	21		Schwingungsauswertung
Candidate	0,08	1	22		Schwingungsbeurteilung
Candidate	0,25	3	21		Schwingungselektronik
Candidate	0,08	1	23		schwingungsempfindliche
Candidate	0,08	1	15		schwingungsfrei

The Work in the Extraction

- Search for specific things, like hyphens, slashes or ampersands.

There might be good candidates.

Status	$	#	Length	Hides	Source
Candidate	0,33	4	27		Anti-Korruptions-Leitlinien
Candidate	0,08	1	27		Grundlast-/Wechselschaltung
Candidate	0,08	1	27		Argon-Sauerstoff-Entkohlung

Status	$	#	Length	Hides	Source
Candidate	0,17	2	17		Plug&Play-Prinzip
Candidate	0,08	1	16		Plug&Play-System

Status	$	#	Length	Hides	Source
Candidate	0,58	6	21		leak-indicating·liquid
Candidate	0,97	10	21		multi-labelling·system

There might a list that is easy to drop.

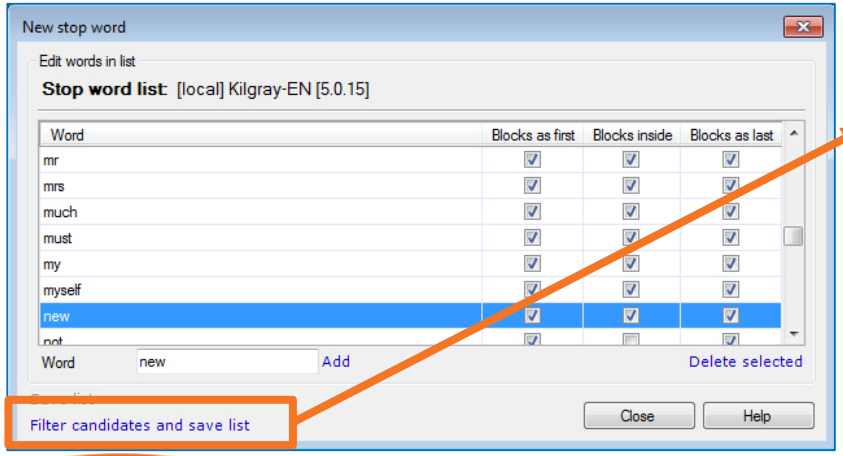
Status	$	#	Length	Hides	Source
Candidate	0,48	5	19		directive·2014/32/EU
Candidate	0,29	3	19		EN·1998-1/NA:2011-01
Candidate	2,52	26	19		directive·2014/34/EU
Candidate	0,39	4	19		directive·2006/42/EG
Candidate	0,68	7	19		directive·[illegible]

Status	$	#	Length	Hides	
Candidate	0,24	7	6		+/-10%
Candidate	0,29	3	7		<10·mg/l
Candidate	0,53	3	10		=·1,000·m³/h
Candidate	0,39	4	4		=·d/2
Candidate	0,71	4	5		=·d/2·+
Candidate	0,29	3	6		≤·90m/s
Candidate	0,48	5	5		≤·f/fn
Candidate	0,89	5	6		≤·f/fn·<

ss@zaac.de

The Work in the Extraction List

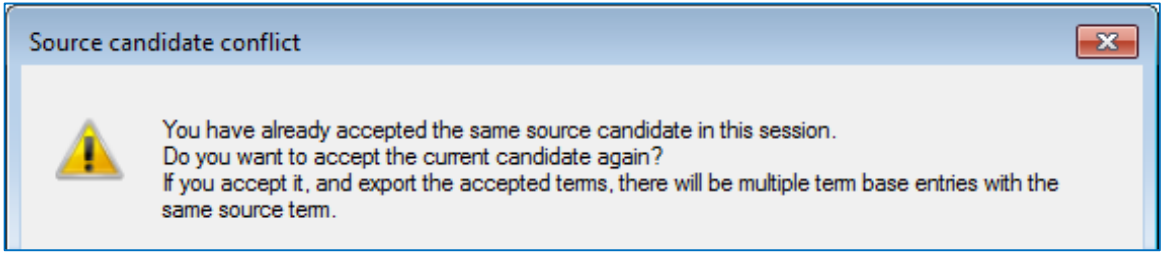
- Send terms to the stop word list…
- …and filter candidates for the stop word to drop them in one go.



Status	S	#	Length	Hides	Source
Candidate	107,92	3153	3		new
Candidate	0,39	4	4		new·E
Candidate	0,68	7	4		•·new
Candidate	3,46	101	4		News
Candidate	0,17	5	4		NEWA
Candidate	0,68	20	4		anew
Candidate	0,29	3	5		=>·new
Candidate	2,88	84	5		newly
Candidate	0,29	3	5		new·VM

The Details

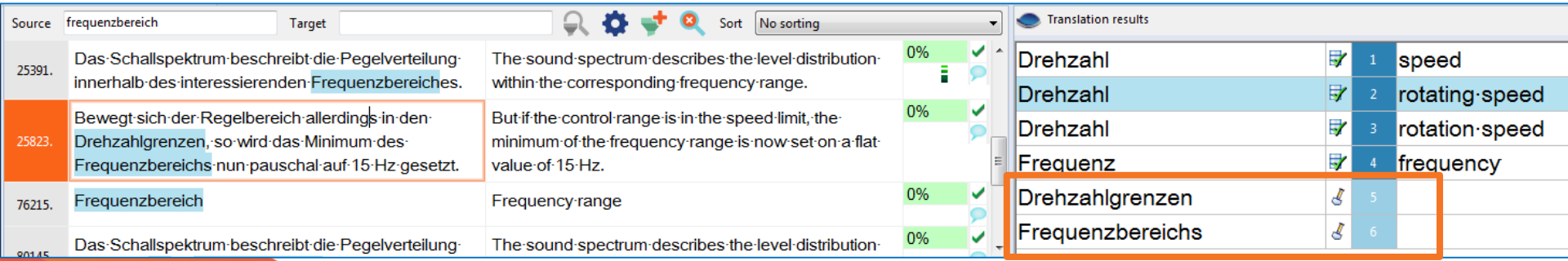
- If you try to accept an already accepted term, memoQ will notify you.



- The status bar shows how many candidates there are, how many have been accepted or dropped already

The Details

- Accepted (confirmed) term candidates appear in the translation results list with a spade symbol (term mining).
- They are used for matching like the alignments from the LiveDocs.



The Numbers

- From my own extractions (EN or DE)
- General language documents:
  - About 3-5% of all words (total word count) are possible terms
- Subject-matter specific documents:
  - About 5-7% of all words are possible terms

OR

- Between 10-20% of the term candidates in the extraction list are possible terms.

The Numbers

- Extraction speed
  - Going through a list of candidates in the extraction and just confirming and dropping source language terms: 2000-3000 candidates / hour
- Number of extracted terms (on average)
  - Source language only: 100+ terms/hour
  - Bilingual extraction: 80+ terms/hour
  - Bilingual with context, comments and other research in-between: 20-30+ terms/hour

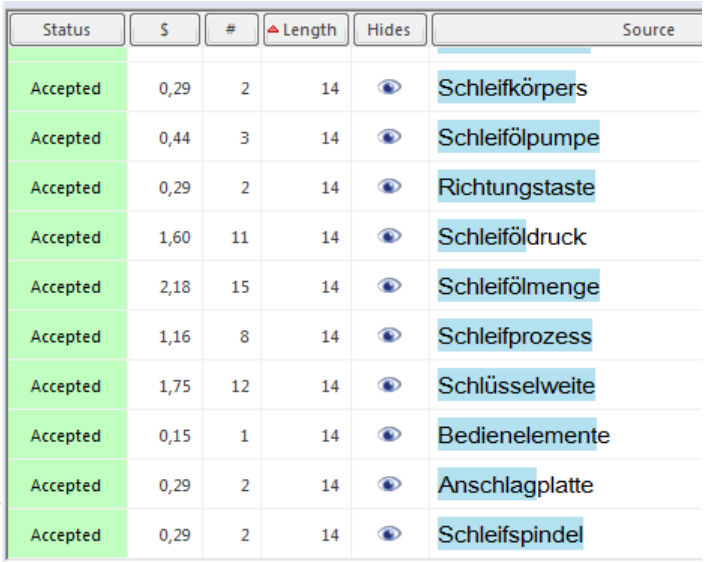
- Additional work
  - Sorting the list
  - Adding comments, instructions, questions…
  - Decide if all terms really are terms to keep

Wishes for the future

- Extraction from server term bases / sending a term extraction job to a translator/terminologist
- Regex to search/filter entries
- Possibility to create a list of endings that can help to select the base form out of singular, plural, genitive form… (language specific)
- Add the feature to mark a possible translation (in green, like the "guess translation" feature in the concordance search)
- Sort by one-word candidates and multi-word candidates
- When showing candidates with known terms only, it would be nice to know how many there are and also to have a list of only candidates that have no known term

Wishes for the future

- Search for terms that consist of capital letters only or have capital letter in the middle (acronyms, product names)
- Being able to sort for candidates that have a full match from the term base, those that have several matches from the term base and those where there are x characters (by length) added to the term base term



Status	$	#	Length	Hides	Source
Accepted	0,29	2	14		Schleifkörpers
Accepted	0,44	3	14		Schleifölpumpe
Accepted	0,29	2	14		Richtungstaste
Accepted	1,60	11	14		Schleiföldruck
Accepted	2,18	15	14		Schleifölmenge
Accepted	1,16	8	14		Schleifprozess
Accepted	1,75	12	14		Schlüsselweite
Accepted	0,15	1	14		Bedienelemente
Accepted	0,29	2	14		Anschlagplatte
Accepted	0,29	2	14		Schleifspindel

zerfass@zaac.de

memoq fest

Term Extraction

29-31 May 2019 I Budapest, Hungary

zerfass@zaac.de