MULTILINGUAL TERMINOLOGY EXTRACTION

TERMVAR WORKSHOP ON TERMINOLOGICAL VARIATION

HILDESHEIM, JULY 2017

AGENDA

- MOTIVATION
- AUTOMATIC TERM EXTRACTION
- MULTILINGUAL TERM EXTRACTION
- SOME RESULTS
- WHAT YOU COULD DO WITH THESE RESULTS

BUT FIRST…

MOST OF THE PRESENTED WORK WAS DONE AS MEMBER OF A WORKING GROUP AT ACROLINX GMBH BERLIN

HTTPS://WWW.ACROLINX.DE/

MOTIVATION

WHAT IS IT GOOD FOR?

WHERE DOES TERMINOLOGY COME FROM?

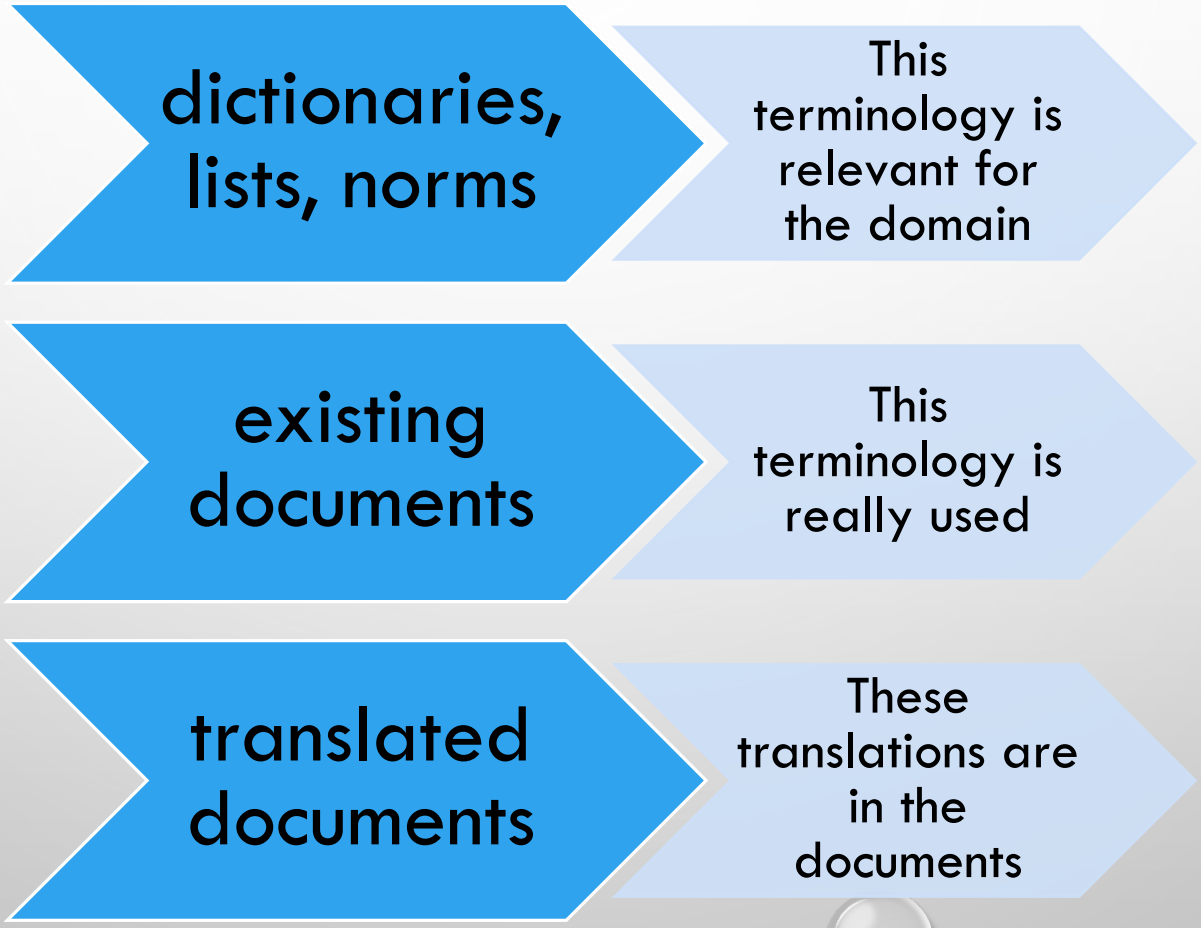
- OLD TERMINOLOGY IN LISTS
- WORD LISTS FROM USER DICTIONARIES
- WORD LISTS OF TRANSLATORS
- PRODUCT LISTS

```
water pressure
overheating
capacity
valve
tube
hot water
program sequence
```

	sales designation	designation
2	G 1020	dish washer G 1020
3	G 1020 i	dish washer G 1020 i
4	G 1020 SC	dish washer G 1020 SC
5	G 1020 SCi	dish washer G 1020 SCi
6	G 1020 SCU	dish washer G 1020
7	G 1020 SCU	dish washer G 1020 SCU
8	G 1022	dish washer G 1022
9	G 1022 i	dish washer G 1022 i
10	G 1022 SC	dish washer G 1022 SC
11	G 1022 SCi	dish washer G 1022 SCi/integrated/without Bl.
12	G 1022 SCU	dish washer G 1022 SCU
13	G 1022 U	dish washer G 1022 U
14	G 1023 i	dish washer G 1023 i

German	English	French	Spanish	Italian	
Wasserdruck	water pressure	pression d'eau	presión hidráulica	pressione dell'ac	
Überhitzung	overheating	surchauffe	recalentamiento	surriscaldamento	
Gläsereinsatz					
Füllmenge	capacity				
Geschirrspülertür					
Besteckaufsatz					
Unterbaugerät					
Klappe	valve	clapet	chapaleta	chiusino	床
Schlauch	tube	boyau	manga	budello	
Warmwasser	not water	eau chaude	agua caliente		
Programmablauf	program sequence	enchaînement	marcha del programa		
Anlagebügel					

WHERE DOES TERMINOLOGY COME FROM?



WE NEED TERMINOLOGY THAT IS…

- … RELEVANT FOR THE FIELD
- … USED IN THE DOCUMENTS
- … USED BY TRANSLATORS

automatic extraction, multilingual

AUTOMATIC TERM EXTRACTION

MONOLINGUAL

STATISTIC INFORMATION FOR TERM EXTRACTION

- EXTRACTION CORPUS
  - TEXTS FROM THE DOMAIN
- REFERENCE CORPUS
  - TEXTS FROM OTHER DOMAINS, SUCH AS NEWSPAPER CORPUS
- STATISTIC COMPARISON OF WORDS IN THESE TEXTS
  1. WORDS THAT OCCUR ONLY IN THE EXTRACTION CORPUS: TERMINOLOGY
  2. WORDS THAT OCCUR MORE FREQUENTLY IN THE EXTRACTION CORPUS: POTENTIAL TERMINOLOGY
  3. WORDS WITH SIMILAR FREQUENCY IN BOTH CORPORA: STOP WORDS
  4. WORDS THAT OCCUR MORE FREQUENTLY IN THE REFERENCE CORPUS: POSSIBLY NO TERMINOLOGY FOR THIS DOMAIN

Heyer, Quasthoff, Wittig 2006: S. 95f.

EXAMPLE TEXT FOR TERM EXTRACTION

Starker Nebel hindert Mücken am Fliegen

Die Kollision mit einem Regentropfen übersteht eine Mücke problemlos, obwohl er rund fünfzig Mal größer ist als das Insekt. Bei starkem Nebel haben Mücken allerdings Probleme mit dem Fliegen. Forscher haben herausgefunden, warum das so ist.

Die zahllosen winzigen Nebeltröpfchen blockieren die Schwingkölbchen der Mücke - zwei kleine, hinter den Flügeln sitzende Lagesensoren. Das haben US-amerikanische Forscher mittels Hochgeschwindigkeitsaufnahmen herausgefunden. Demnach kollidieren die schwingenden Sensoren in jeder Sekunde mit tausenden Nebeltröpfchen und funktionieren dadurch nicht mehr richtig. Als Folge könne die Mücke ihre Körperposition im Flug nicht mehr ermitteln und verliere ihre stabile Fluglage, berichten die Wissenschaftler am Montag auf einer Physiker-Tagung im kalifornischen San Diego.

„Moskitos sind auch im Regen gute Flieger, aber bei Nebel gelingt ihnen dies nicht“, schreiben Andrew Dickerson und seine Kollegen vom Georgia Institute of Technology. Ein Regentropfen sei rund 50 Mal so groß wie eine Mücke, ein Zusammenstoß daher vergleichbar der Kollision eines Menschen mit einem Bus. Obwohl das winzige Insekt bei einem Regenguss im Durchschnitt alle 20 Sekunden mit einem Tropfen kollidiere, überstehe es dies schadlos und fliege weiter.

„Starker Nebel besteht dagegen aus Tröpfchen, die 100 Mal kleiner sind als bei Regen, dennoch kann die Mücke unter diesen Bedingungen nicht mehr fliegen“, schreiben die Forscher. Ähnlich wie moderne Flugzeuge müsse das Insekt bei starkem Nebel am Boden bleiben. Warum, sei bisher unklar gewesen.

Frankfurter Rundschau: http://www.fr-online.de/natur/muecke-starker-nebel-hindert-muecken-am-fliegen,5028038,20907952.html

IMPLEMENTATION: STATISTIC ANALYSIS

- EXTRACTION CORPUS: TEXT ABOUT MOSQUITOS
- REFERENCE CORPUS: 1,000 MOST FREQUENT WORDS OF GERMAN (WORTSCHATZ LEIPZIG)
- Words only in extraction corpus:
  - ['winzige', 'fünfzig', 'fliege', 'blockieren', 'menschen', 'boden', 'größer', 'sitzende', 'insekt', 'unklar', 'verliere', 'institute', 'wissenschaftler', 'schwingkölbchen', 'bedingungen', 'gelingt', 'demnach', 'physiker-tagung', 'folge', 'of', 'durchschnitt', 'mücke', 'fliegen', 'problemlos', 'diego', 'san', 'flügeln', 'andrew', 'flug', 'hindert', 'tausenden', 'mittels', 'bus', 'berichten', '-', 'überstehe', 'sekunde', 'winzigen', 'vergleichbar', 'kollegen', 'flugzeuge', 'sensoren', 'probleme', 'mücken', 'lagesensoren', 'regenguss', 'schwingenden', 'übersteht', 'ähnlich', 'us-amerikanische', 'kalifornischen', 'zusammenstoß', 'sekunden', 'nebeltröpfchen', 'starkem', 'montag', 'funktionieren', 'zahllosen', 'flieger', 'körperposition', 'moskitos', 'starker', 'kollidieren', 'herausgefunden', 'kleiner', 'ermitteln', 'regen', 'stabile', 'nebel', 'forscher', 'schadlos', 'regentropfen', 'hochgeschwindigkeitsaufnahmen', 'fluglage', 'schreiben', 'dickerson', 'kollidiere', 'tropfen', 'georgia', 'tröpfchen', 'moderne', 'technology', 'kollision']

MINIMAL LINGUISTIC INFORMATION FOR TERM EXTRACTION

- EXTRACT ALL CAPITAL WORDS IN A GERMAN TEXT
- IF TWO CAPITAL WORDS OCCUR IN A ROW, ASSUME THIS IS A MULTIWORD EXPRESSION
- BUT NOT THE FIRST WORD IN A SENTENCE
- EXCEPT IF IT IS A NOUN

RESULTS OF EXTRACTION WITH MINIMAL LINGUISTIC INFORMATION

['Andrew Dickerson', 'Boden', 'Bedingungen', 'Bus', 'Durchschnitt', 'Dickerson', 'Diego', 'Forscher', 'Flügeln', 'Flieger', 'Folge', 'Flugzeuge', 'Fluglage', 'Flug', 'Fliegen', 'Georgia Institute', 'Hochgeschwindigkeitsaufnahmen', 'Institute', 'Insekt', 'Körperposition', 'Kollision', 'Kollegen', 'Lagesensoren', 'Mal', 'Menschen', 'Mücke', 'Montag', 'Mücken', 'Nebel', 'Nebeltröpfchen', 'Probleme', 'Physiker-Tagung', 'Regentropfen', 'Regenguss', 'Regen', 'Sekunden', 'San Diego', 'Sekunde', 'Schwingkölbchen', 'Sensoren', 'Tropfen', 'Tröpfchen', 'Technology', 'US-amerikanische Forscher', 'Wissenschaftler', 'Zusammenstoß']

LINGUISTIC INFORMATION FOR TERM EXTRACTION

- SYNTACTIC CATEGORIES (NOUNS, VERBS, ADJECTIVES,…)
- COMPOUND ANALYSIS
- ACRONYMS
- PRODUCT NAMES (NAMED-ENTITY RECOGNITION)
- SPELLING
- NEW WORDS

PATTERNS WITH LINGUISTIC INFORMATION

- PATTERNS WITH LINGUISTIC INFORMATION DETECT POTENTIAL TERMINOLOGY
  - NOUN, SUCH AS *TEXTFELD*
  - MULTIWORD EXPRESSION, SUCH AS *SERVICE CENTER*
  - COMPOSED ADJECTIVES, SUCH AS *EXTRAKTIONS-SPEZIFISCH*
  - ABBREVIATIONS, SUCH AS *STT*
  - ACRONYM WITH DEFINITION, SUCH AS *NETWARE LOADABLE MODULE (NLM)*
  - TRADE MARK, SUCH AS *TERM HARVESTING*™
  - CAMEL CASE, SUCH AS *NFINIBAND*
  - WORD WITH SLASH, SUCH AS *KALTWASSER/WARMWASSER*
  - COMPANY NAME, SUCH AS *ABC GMBH*

ACROLINX EXTRACTION

- COMBINING STATISTIC AND LINGUISTIC INFORMATION
- OBSERVING TERMINOLOGY ALREADY EXTRACTED
- OBSERVING (POTENTIAL) SPELLING ERRORS

RESULTS OF COMPLEX TERM EXTRACTION WITH ACROLINX

Flieger	Moskitos sind auch im Regen gute Flieger, aber bei Nebel gelingt ihnen dies nicht , schreiben Andrew Dickerson und seine Kollegen vom Georgia Institute of Technology.	NN	1
Fluglage	Als Folge könne die Mücke ihre Körperposition im Flug nicht mehr ermitteln und verliere ihre stabile Fluglage, berichten die Wissenschaftler am Montag auf einer Physiker-Tagung im kalifornischen San Diego.	NN	1
Hochgeschwindigkeitsaufnahmen	Das haben US-amerikanische Forscher mittels Hochgeschwindigkeitsaufnahmen herausgefunden.	NN	1
Körperposition	Als Folge könne die Mücke ihre Körperposition im Flug nicht mehr ermitteln und verliere ihre stabile Fluglage, berichten die Wissenschaftler am Montag auf einer Physiker-Tagung im kalifornischen San Diego.	NN	1
Kollision	Die Kollision mit einem Regentropfen übersteht eine Mücke problemlos, obwohl er rund fünfzig Mal größer ist als das Insekt. Ein Regentropfen sei rund 50 Mal so groß wie eine Mücke, ein Zusammenstoß daher vergleichbar der Kollision eines Menschen mit einem Bus.	NN	2
Lagesensoren	Die zahllosen winzigen Nebeltröpfchen blockieren die Schwingkölbchen der Mücke - zwei kleine, hinter den Flügeln sitzende Lagesensoren.	NN	1
Moskitos	Moskitos sind auch im Regen gute Flieger, aber bei Nebel gelingt ihnen dies nicht , schreiben Andrew Dickerson und seine Kollegen vom Georgia Institute of Technology.	NN	1

Mücken	Starker Nebel hindert Mücken am Fliegen Die zahllosen winzigen Nebeltröpfchen blockieren die Schwingkölbchen der Mücke - zwei kleine, hinter den Flügeln sitzende Lagesensoren. …	NN	7
Nebeltröpfchen	Die zahllosen winzigen Nebeltröpfchen blockieren die Schwingkölbchen der Mücke - zwei kleine, hinter den Flügeln sitzende Lagesensoren. …	NN	2
Physiker-Tagung	… berichten die Wissenschaftler am Montag auf einer Physiker-Tagung im kalifornischen San Diego.	NN	1
Regenguss	Obwohl das winzige Insekt bei einem Regenguss im Durchschnitt alle 20 Sekunden mit einem Tropfen kollidiere, überstehe es dies schadlos und fliege weiter.	NN	1
Regentropfen	Die Kollision mit einem Regentropfen übersteht eine Mücke problemlos, obwohl er rund fünfzig Mal größer ist als das Insekt. …	NN	2
Sensoren	Demnach kollidieren die schwingenden Sensoren in jeder Sekunde mit tausenden Nebeltröpfchen und funktionieren dadurch nicht mehr richtig.	NN	1
Tröpfchen	Starker Nebel besteht dagegen aus Tröpfchen, die 100 Mal kleiner sind als bei Regen, dennoch kann die Mücke unter diesen Bedingungen nicht mehr fliegen , schreiben die Forscher.	NN	1
US-amerikanische	Das haben US-amerikanische Forscher mittels Hochgeschwindigkeitsaufnahmen herausgefunden.	ADJA	1
Zusammenstoß	Ein Regentropfen sei rund 50 Mal so groß wie eine Mücke, ein Zusammenstoß daher vergleichbar der Kollision eines Menschen mit einem Bus.	NN	1

MULTILINGUAL TERM EXTRACTION

ADDING SOME MACHINE TRANSLATION TECHNOLOGY

WHAT WE NEED FOR MULTILINGUAL TERM EXTRACTION

- A MONOLINGUAL EXTRACTION SYSTEM FOR ONE OF THE INVOLVED LANGUAGES
  - SUCH AS ACROLINX
- A LARGE CORPUS OF ALIGNED SENTENCES
  - SUCH AS A TRANSLATION MEMORY
- A STATISTICAL MACHINE TRANSLATION SYSTEM
  - SUCH AS MOSES

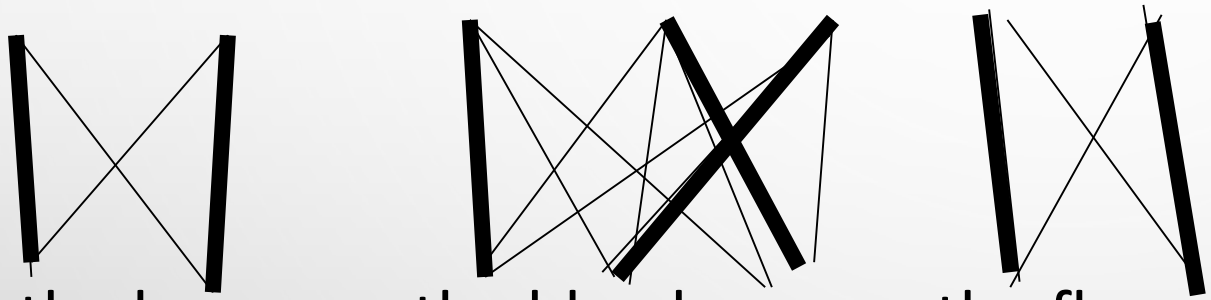
FROM ALIGNED SENTENCES TO ALIGNED WORDS

- IN A LARGE CORPUS OF ALIGNED SENTENCES
  - COUNT THE OCCURRENCE OF A WORD IN THE SOURCE LANGUAGE (SL) SENTENCES AND A WORD IN THE TARGET LANGUAGE (TL) SENTENCES
  - TAKE THE POSITION OF WORDS IN A SENTENCE INTO ACCOUNT
  - TAKE THE SENTENCE LENGTH INTO ACCOUNT

  → THE PROBABILITY THAT A SL WORD CORRESPONDS TO A TL WORD

LEARNING TRANSLATIONS FROM PARALLEL DATA

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

- model learns: *la* is often aligned with *the*
- model learns: *maison* is often aligned with *house*
- it becomes clear: *fleur* is *flower, bleu* is *blue*

Phillip Koehn, EMNLP, 2007

STATISTICS ON WORDS

- TRANSLATIONS OF GERMAN „HAUS“ IN A DICTIONARY:
  - HAUS — HOUSE, BUILDING, HOME, HOUSEHOLD, SHELL
- SOME OF THESE TRANSLATIONS OCCUR MORE FREQUENTLY THAN OTHERS
- COUNTS AFTER SEARCHING IN A CORPUS OF ALIGNED SENTENCES:

Translation of Haus	Count
house	8,000
building	1,600
home	200
household	150
shell	50

Phillip Koehn, EMNLP, 2007

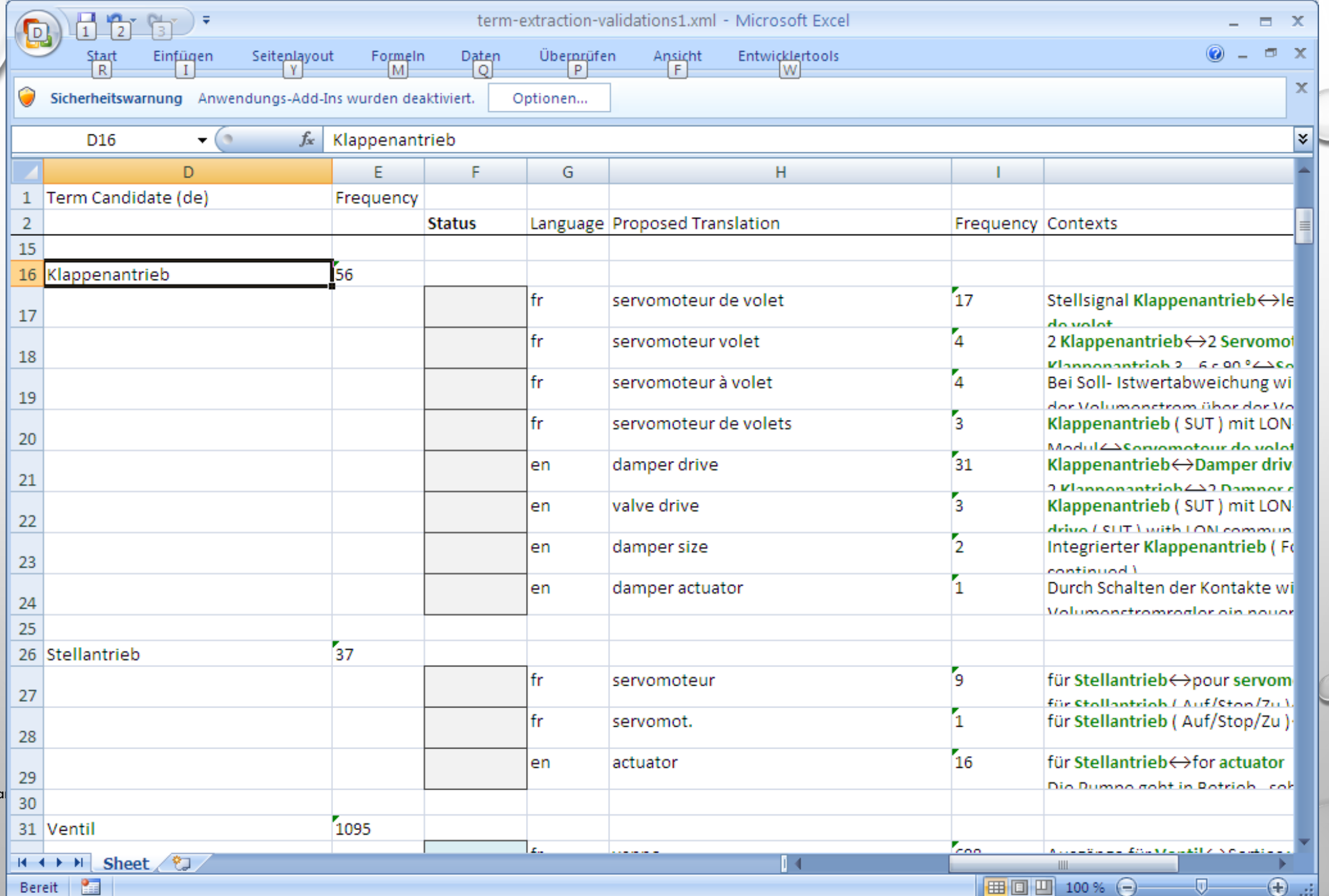
MAXIMUM LIKELIHOOD ESTIMATION

$$PF(e) = \begin{cases} 0.8 \text{ if } e = \text{house}, \\ 0.16 \text{ if } e = \text{building}, \\ 0.02 \text{ if } e = \text{home}, \\ 0.015 \text{ if } e = \text{household}, \\ 0.005 \text{ if } e = \text{shell}. \end{cases}$$

PROCESS OF MULTILINGUAL TERMINOLOGY EXTRACTION

SOME RESULTS

UNEXPECTED INSIGHTS TO TRANSLATION MEMORIES



	D	E	F	G	H	I	
1	Term Candidate (de)	Frequency					
2			**Status**	Language	Proposed Translation	Frequency	Contexts
15							
16	Klappenantrieb	56					
17				fr	servomoteur de volet	17	Stellsignal **Klappenantrieb**↔le
18				fr	servomoteur volet	4	2 **Klappenantrieb**↔2 **Servomot**
19				fr	servomoteur à volet	4	Bei Soll- Istwertabweichung wi
20				fr	servomoteur de volets	3	**Klappenantrieb** (SUT) mit LON
21				en	damper drive	31	**Klappenantrieb**↔**Damper driv**
22				en	valve drive	3	**Klappenantrieb** (SUT) mit LON
23				en	damper size	2	Integrierter **Klappenantrieb** (F
24				en	damper actuator	1	Durch Schalten der Kontakte wi
25							
26	Stellantrieb	37					
27				fr	servomoteur	9	für **Stellantrieb**↔pour **servom**
28				fr	servomot.	1	für **Stellantrieb** (Auf/Stop/Zu)
29				en	actuator	16	für **Stellantrieb**↔for **actuator**
30							
31	Ventil	1095					

- Hubtisch
  - lift table
  - table lift
  - lifting platform
- Grundstellung
  - starting position
  - basic position
  - home position
  - basic setting
  - initial position
  - starting pos.
  - normal position

- adaptor
  - Zwischenstück
  - Adaptor
- operating unit
  - Bedieneinheit
  - Bediengerät

	A	B	E
1	Term Candidate	Frequency	
2		Translation	Score
415	Tankinhalt	12	
416		tank volume	5.769225
417		amount of fuel in the tank	4.640156
418		tank content	4.040953
419		tank reservoir	1.04167
420		amount of fuel	0.850697
421	Spurweite	12	
422		track width	13.57089
423		tread width	7.021452
424		tread	2.165127
425		wheel tread	1.5444
426	Aufstiegsleiter	12	
427		access ladder	28.60122
428		ladder	11.68565
429	Befüllung	11	
430		filling	4.869595
431		filling process	2.926828
432	Störungssuche	11	
433		Troubleshooting	38.8724
434	Spritzpumpe	11	
435		spraying pump	66.54593
436	Luftdruck	11	
437		air pressure	36.80594
438	Sollwert	11	
439		desired value	9.612736
440		target rate	5.142257
441		set point	3.200166
442		target	3.118112

WHAT YOU COULD DO WITH THESE RESULTS

check and improve your TMX	• clusters in source language and target languages
improve the quality of your (human) translations	• provide a bilingual dictionary for translators • provide a checking method for terminology in the target languages • check translations for inconsistencies in terminology translation
improve the quality of your machine translation results	• improve the dictionary for MT • provide terminology for target language checking • improve the quality of training data
improve the quality of your texts	• find term clusters, synonyms and variants on the source language text
improve your SEO	• find synonyms

THANK YOU!

MELANIE SIEGEL

HOCHSCHULE DARMSTADT

INFORMATIONSWISSENSCHAFT

MELANIE.SIEGEL@H-DA.DE

WWW.MELANIESIEGEL.DE