



**University of  
Zurich** <sup>UZH</sup>

# Termextraktion & Fremdsprachenlernen

CAS in Translation Technology and AI

---

Martin Volk & Johannes Graën

5. November 2021

# Termextraktion mittels Wortalignment

---

## MultiTerm Extract

*„MultiTerm Extract checks the frequency of terms at a sub-segment level and enables you to build project glossaries without having to manually search for the terms. You can automatically locate and extract potential monolingual or bilingual terms from existing documentation or TMs to build termbases and glossaries quickly.“*

*„SDL MultiTerm Extract uses a statistical extraction method to determine the frequency of the appearance of candidate terms. It extracts term candidates and, for multilingual termbases, their probable translations found in sentences, incomplete sentence fragments and strings of code. The extracted terms are presented as a term candidate word or phrase.“*

## projectTermExtract

*„This projectTermExtract plugin adds a very neat feature to Studio allowing you to extract term candidates from your Project, or specific files within a project.“*

# Termextraktion – Trados (2)

Project Terms Cloud

ProjectTermExtract

Project term settings

Blacklist:

Enter term to blacklist:

Use regular expressions

Add

Delete

Reset

Load

Save

Terms occurrences:

Terms length:

Extract Terms From Selected Files

Name	Words	Status	Progress	Size	Usage	File Type	Iden.
1880 Horse carriage dock...	101	In Translat...	9%	217 KB	Translatable	Wordprocessing...	...
1896 Stanley Pumbout do...	115	In Translat...	10%	186 KB	Translatable	Wordprocessing...	...
1902 Mercedes-Simplex 6...	431	In Translat...	9%	503 KB	Translatable	Wordprocessing...	...
1910 Austro-Daimler 22 B...	458	In Translat...	6%	482 KB	Translatable	Wordprocessing...	...
1929 Duesenberg J dock s...	502	In Translat...	23%	434 KB	Translatable	Wordprocessing...	...
1932 Duesenberg SJ dock...	464	In Translat...	13%	511 KB	Translatable	Wordprocessing...	...
1943 Jaguar XK 120 Road...	525	In Translat...	17%	505 KB	Translatable	Wordprocessing...	...
1959 Aston Martin D64 G...	521	In Translat...	14%	439 KB	Translatable	Wordprocessing...	...

*„memoQ processes the text and gives you a list of candidates - possible terms. There may be a lot of garbage in the list: You may need to clean it up, filter, and edit it - and confirm “true,, terms before you can add those to a term base. After the extraction runs, memoQ opens the candidate list editor where you can do all this.“*

# Termextraktion – memoQ (2)

Extract candidates



Session name

GG\_TEX\_test

Sources

Translation documents

Every document

Selected documents

Translation memories

All memories in project

Primary TM

Selected TMs

LiveDocs corpus documents

All documents shown

Selected documents

Options

General

Maximum length (words)

4

Minimum frequency

3

Expression delimiters

:""0[]?!\«»« »'„”

Length factor

1.50

Ignore words with numbers

Single-word terms

Minimum length (characters)

3

Minimum frequency

3

Term base lookup

Look up candidates

All term bases in project

Term base with the highest rank only

Stop words

Stop word list

[local] Kilgray-EN [5.0.15]

Save as...

Word	Blocks as first	Blocks inside	Blocks as last
a	<input checked="" type="checkbox"/>	<input 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"/>
across	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
after	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Word

Add

Delete selected

OK

Cancel

Help

*„Gleichzeitig unterstützen MindReader und Transit die Terminologearbeit, z.B. mit der komfortablen Terminologieextraktion aus dem Übersetzungsprojekt oder dynamisch generierten Verwendungsbeispielen aus dem Translation Memory.“*

*„Mit dem Translation-Memory-System Transit NXT kann der Nutzer auch Terminologieextraktion durchführen. Möglich ist die einsprachige Terminologieextraktion aus einem Ausgangstext.“*

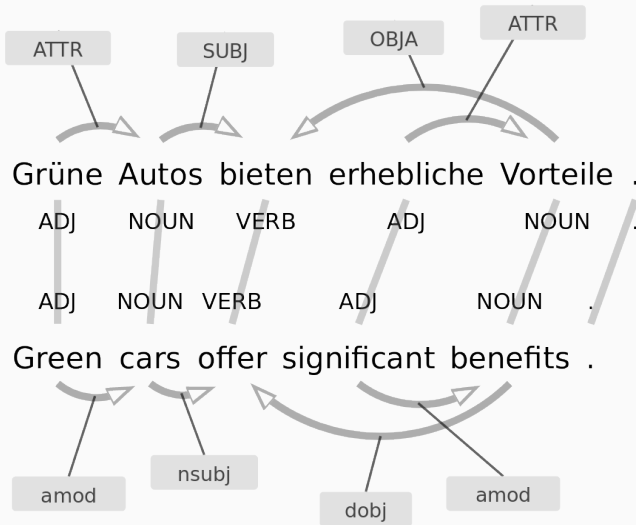
*„Consistent terminology is a key precondition for high-quality translations. For this purpose, apart from the terminology management crossTerm, Across offers the possibility of extracting potential terms (term candidates) and subsequently determining and translating the actual terms before starting to translate the source document.“*



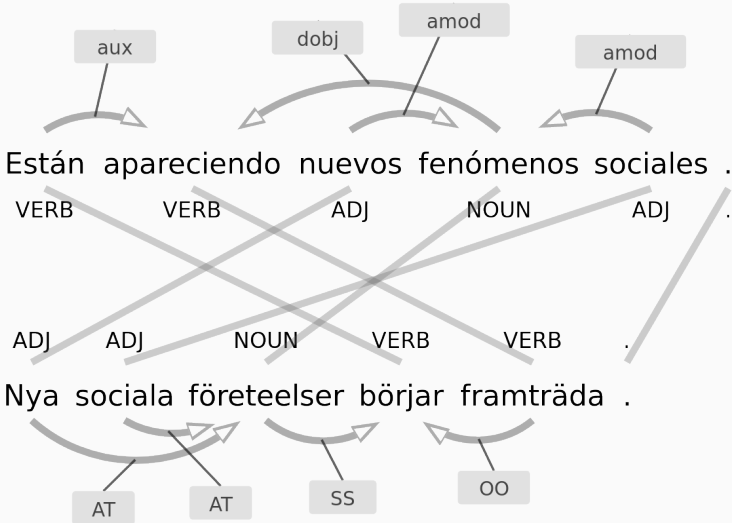
- Terme werden meist einsprachig ermittelt
- Die Methoden werden nicht offengelegt («a statistical extraction method»)
- Die Benutzer müssen Termkandidaten manuell einer Termdatenbank hinzufügen
- Alignierung scheint bei keinem der Produkte in Verwendung zu sein

- 1:n-Alignierungen als Indikator («nach wie vor»)
- Hinzuziehen syntaktischer Strukturen (ParZu →)
- Abschätzen der Erwartungshaltung vor oder nach einer Folge von Wörtern («formulaic speech»)
  - wie kann «Bausch und» fortgesetzt werden?
  - was kann vor «oder lebendig» stehen?
  - Beispiel: Binomiale Adverbien →

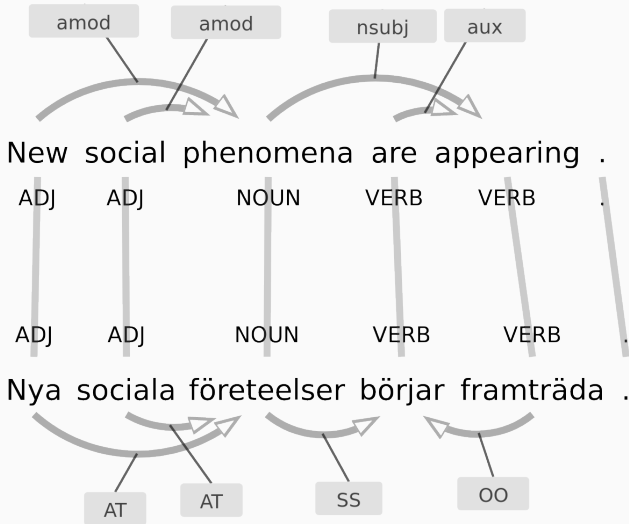
# Alignierung & Syntax (1)



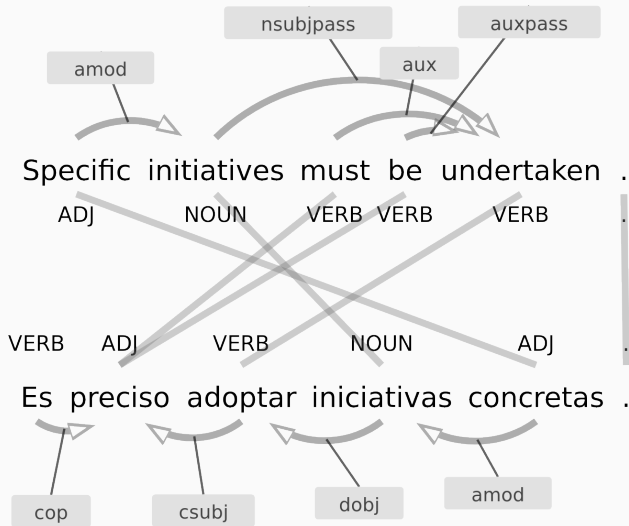
# Alignierung & Syntax (2)



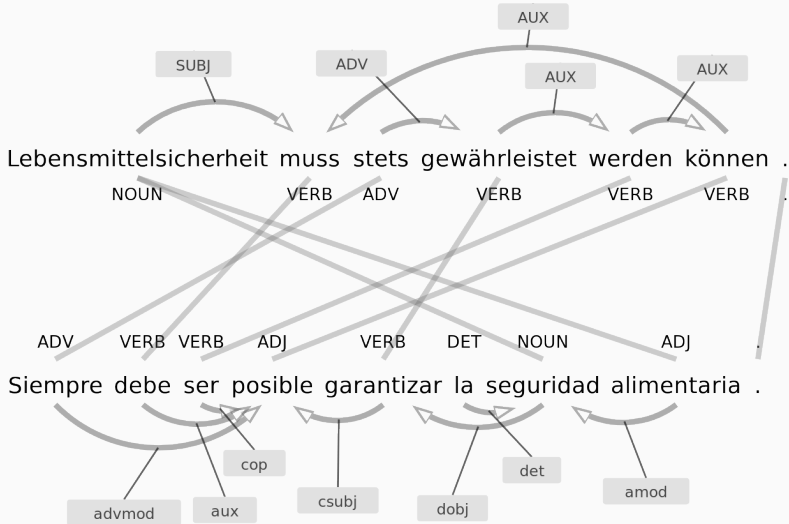
# Alignierung & Syntax (3)



## Alignierung & Syntax (4)



# Alignierung & Syntax (5)



# Sprachenlernen mithilfe von Übersetzungen

---



- authentische Texte
- (oft) professionelle Übersetzungen
- für unterschiedliche Domänen, Texttypen, Register verfügbar
- auch transkribierte Redebeiträge (Parlamentsdebatten, Untertitel) möglich
- Korpusabfragetools für Linguisten wenig geeignet für Lerner

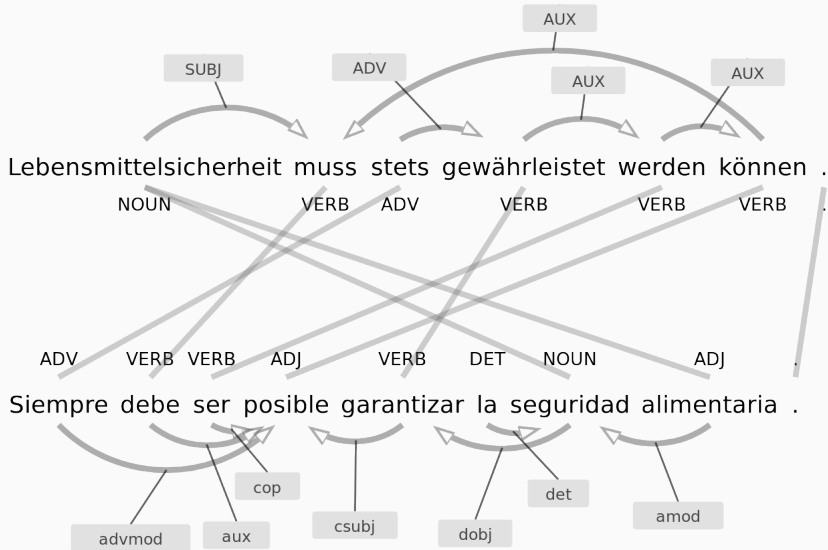
- Lerner explorieren die Zielsprache anhand realer Beispiele (Sätze)
- verschiedene Typen von Parallelität für verschiedene Lernerlevel
- ‹Good Learner Examples› (in Anlehnung an ‹Good Dictionary Examples› (GDEX))

I have also sent a message to the Spanish Government .



También he comunicado a las autoridades españolas .

## Parallelität (2)



## PaCLE

(Parallel Corpora for Language Learning Exercises)

Korpussuchwerkzeug, zum Auffinden paralleler Sätze

einziges Korpus zur Zeit: OpenSubtitles

für Sprachlehrer und -lerner

Nutzer können Satzpaare bewerten und ändern (Crowdsourcing)

einfache Suche mittels regulären Ausdrücken

komplexe Suche über PoS-Tags, syntaktische Beziehungen und  
Alignierungen von Token und Phrasen (future)

visuelle Übersicht über die gefundenen Varianten (future)

automatische Generierung von Übungen (future)

siehe Handout

med tiden

kort och gott

över förväntan

öga mot öga

huvud på ett fat

med flit

sakta men säkert