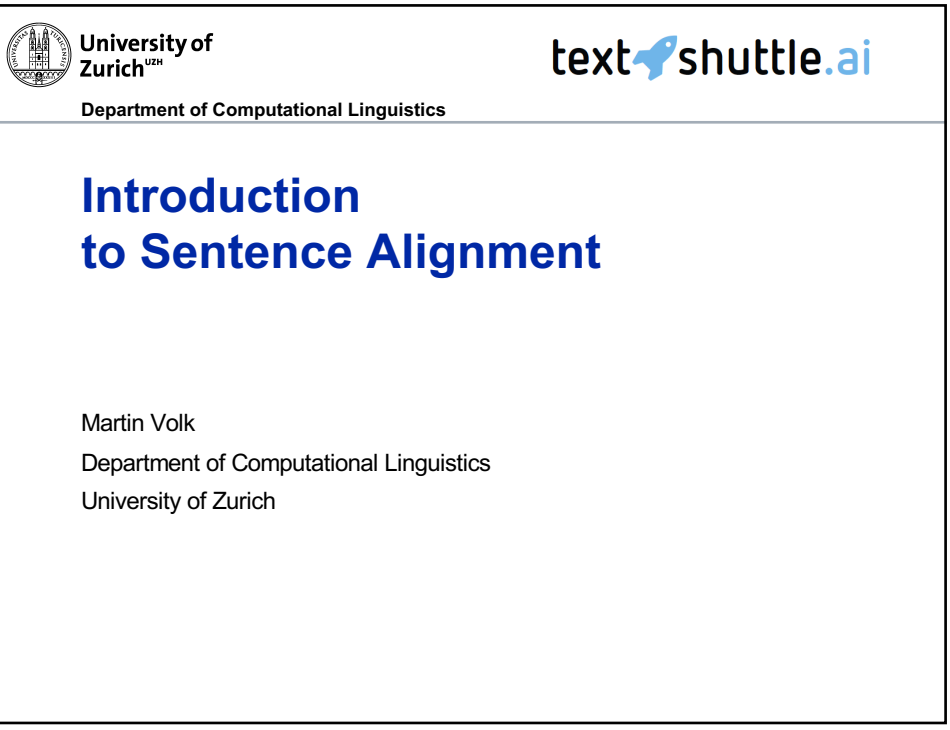
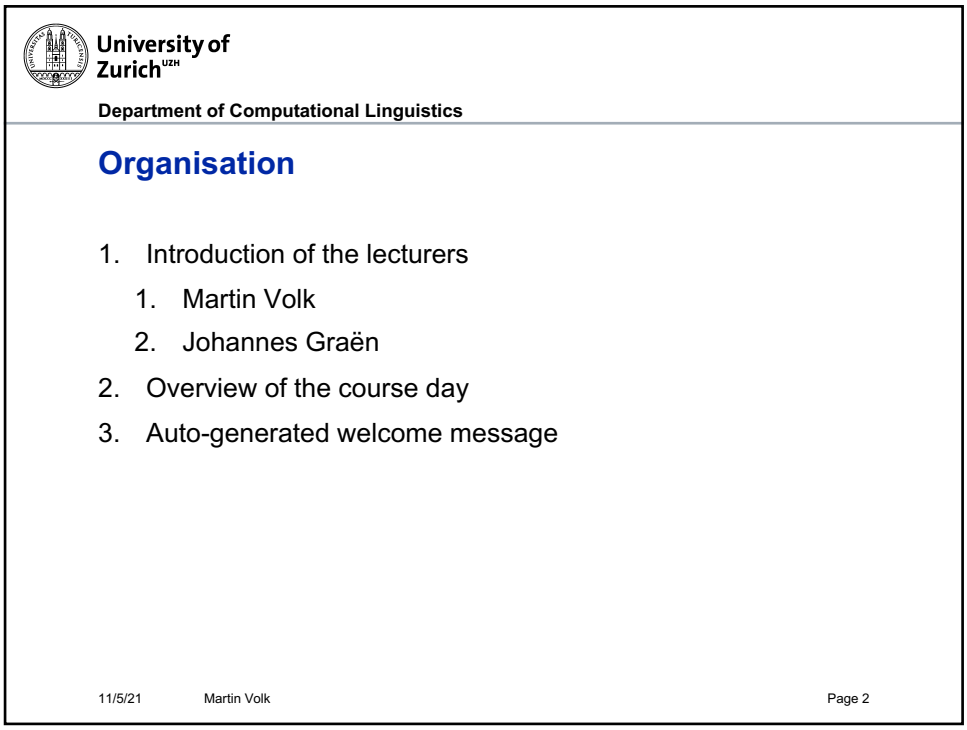

University of Zurich[UZH]
Department of Computational Linguistics
textshuttle.ai
Introduction
to Sentence Alignment
Martin Volk
Department of Computational Linguistics
University of Zurich


University of Zurich[UZH]
Department of Computational Linguistics
Organisation
1. Introduction of the lecturers
1. Martin Volk
2. Johannes Graën
2. Overview of the course day
3. Auto-generated welcome message
11/5/21
Martin Volk
Page 2

Translation Memory

English	German
What's that all about?	Was soll denn das?
I can't wait for this day to be over.	Ich bin so froh, wenn dieser Tag vorbei ist.
Thank God it's only once a year.	Gott sei Dank ist er nur einmal im Jahr.
Maybe I'll meet him tomorrow.	Vielleicht werde ich ihn morgen treffen.
Chris wants to start up a band.	Chris möchte eine Band gründen.
…	…
…	…

University of Zurich[UZH]

Department of Computational Linguistics

Why do translators need "alignment"?

1. to fill Translation Memories
2. to search translated texts
3. to extract bilingual terminology
4. to understand Machine Translation

Why do Computational Linguists need "alignment"?

- for cross-language information transfer
  - e.g. for automatic word sense disambiguation
  - e.g. for proper name recognition
  - e.g. lexicography

University of Zurich[UZH]

Department of Computational Linguistics

InterText Sentence Alignment Demo

Texts from Credit Suisse Bulletin (2003-2014)

- 200 articles in both English and German (tokenized)
  - 13336 segments in English
  - 11835 segments in German

University of Zurich[UZH]

Department of Computational Linguistics

Automatic alignment

… is the process of finding and storing semantically equivalent text units. In other words, text units that are (close to) equal in meaning.

- The cross-language alignment between translated texts is a special case of alignment.
  - We distinguish
    - document alignment
    - **sentence alignment**
    - **word alignment**
- Other alignment examples are …

Automatic alignment

Other alignment cases (to be discussed later today)

1. alignment between different versions of the same text
2. alignment between scan image (of a manuscript) and transcription
3. alignment between audio and transcription

University of Zurich[UZH]
Department of Computational Linguistics

Definition by SDL/Trados

https://docs.rws.com/813470/673714/trados-studio-2021-sr1/aligning-files

Alignment is a mechanism for re-using previous translations. The alignment tool in Trados Studio parses previously translated files into translation units (TUs) so you can add them to a translation memory (TM).

Trados Studio Alignment Mechanism

- "The tool pairs up sentences by looking at the file structure."
- "Trados Studio then analyzes the two sections by unique **features like proper names, numbers, dates, measurements** etc."

University of Zurich[UZH]
Department of Computational Linguistics

Types of Alignment in Trados Studio

1. Align with Review
   - suitable when translations are "not exact"
   - allows major edits
2. Align without Review
3. Align Single File Pair
4. Align Multiple Files
5. Align Retrofit Files (???)

MemoQ

https://docs.memoq.com/ggl-tst/Things/things-alignment.html

"Alignment is a way to reuse previous translations. It means that the original document and its translation are both segmented into translation units, and then the corresponding translation units are matched using **statistical and linguistic algorithms**.

Being a complex process, it may take several minutes, depending on the length of the aligned documents.

Although memoQ's automatic alignment is quite accurate, human revision is necessary for good results." ☺

University of Zurich[UZH]

Department of Computational Linguistics

Parallel Corpora

… are corpora of translated texts.

Where do we find parallel corpora?

University of Zurich[UZH]

Department of Computational Linguistics

Sources of Parallel Corpora

1. Opus
2. Clarin
3. ZH-Parallel Corpus Collection

University of Zurich[UZH]

Department of Computational Linguistics

OPUS – An open source parallel corpus

collected by Jörg Tiedemann, Lars Nygaard

- https://opus.nlpl.eu/
- Europarl
- European Constitution
- Software Manuals
  - Open Office
  - KDE
- Open Subtitles
- …

University of Zurich[UZH]

Department of Computational Linguistics

Parallel Corpora - Examples

The Europarl corpus
– European Parliament proceedings.

The United Nations corpus
– UN texts in 6 languages.

The UZH Text+Berg parallel corpus
– Alpine texts in 2 languages (DE-FR)
– Size: ~ 5 million words / language

Europarl 1996 – 2012

Information taken from http://www.statmt.org/europarl/ (Dec. 2014)

Language	Sentences	Words
Bulgarian	411,636	-
Czech	668,595	13,195,311
Danish	2,323,099	47,761,381
German	2,176,537	47,236,849
Greek	1,517,141	-
English	2,218,201	53,974,751
Spanish	2,123,835	54,806,927
Estonian	692,210	11,358,009
Finnish	2,119,515	33,708,706
French	2,190,579	54,202,850
Hungarian	658,824	12,606,986
Italian	2,081,669	50,259,169
Lithuanian	678,665	11,512,131
Latvian	666,026	12,085,228
Dutch	2,333,816	53,487,257
Polish	387,490	7,087,016
Portuguese	2,121,889	52,300,149
Romanian	402,904	9,663,544
Slovak	674,359	13,116,301
Slovene	634,488	12,665,974
Swedish	2,241,386	45,665,947

17

University of Zurich[UZH]

Department of Computational Linguistics

Europarl Examples DE-EN
(from an EU parliament session in January 2000)

"27"> All dies entspricht den Grundsätzen, die wir stets verteidigt haben.

"26"> This is all in accordance with the principles that we have always upheld.

==========

"28"> Vielen Dank, Herr Segni, das will ich gerne tun.

"27"> Thank you, Mr Segni, I shall do so gladly.

==========

"10"> Wie Sie sicher aus der Presse und dem Fernsehen wissen , gab es in Sri Lanka mehrere Bombenexplosionen mit zahlreichen Toten.

"9"> You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka.

University of Zurich[UZH]

Department of Computational Linguistics

United Nations Corpus

Language	English	French	Spanish	Arabic	Russian	Chinese	German
Documents	96240	85651	70509	65156	77061	65022	3763
Sentences	17098695	14805529	13052875	11050313	13852535	10839473	232225
Words[2]	385894793	377242310	352460926	237412090	278606813	756108566	5848668

= chars

Table 1: Sizes of monolingual data

	fr	es	ar	ru	zh	de
en	96240	68314	63257	74053	62815	3643
fr		68014	63193	73973	62738	3632
es			63241	64230	62707	3632
ar				63194	63031	3677
ru					62842	3635
zh						3886

Table 2: Number of document pairs for the language pairs

Numbers taken from
Eisele and Chen: MultiUN: A Multilingual Corpus from United Nation Documents. LREC, 2010

19

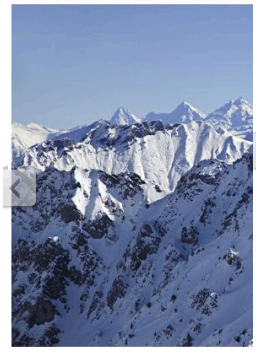
University of Zurich[UZH]

Tourentipp

Rund um den Felsturm

Aussichtsreiche Rundtour einen

In Leysin haben viele Touristen nur Auge
Dabei ist auch der Tour de Famelon ein s
auf Ski führt auf den Gipfel.
Von Stéphane Maire

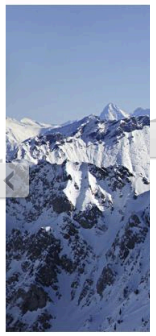


Proposte di itinerari

Attorno alla Tou

Giro panoramico a d

I turisti di passaggio a Le
per gli escursionisti con g
giro.
Del Stéphane Maire



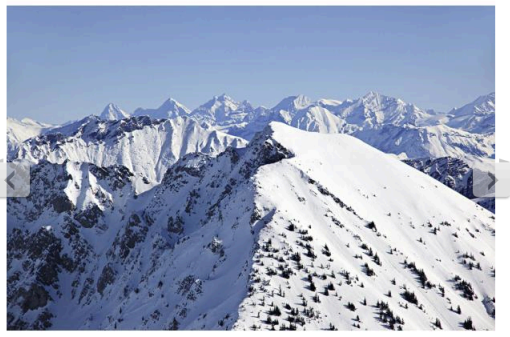
Verso O, la vista raggiunge i giganti dell'Oberland bernese: Eiger, Mönch, Jungfrau, Blüemlisalp e Doldenhorn (da sinistra). In primo piano, il Mont d'Or. © Stéphane Maire

Suggestion de course

Autour de la Tour

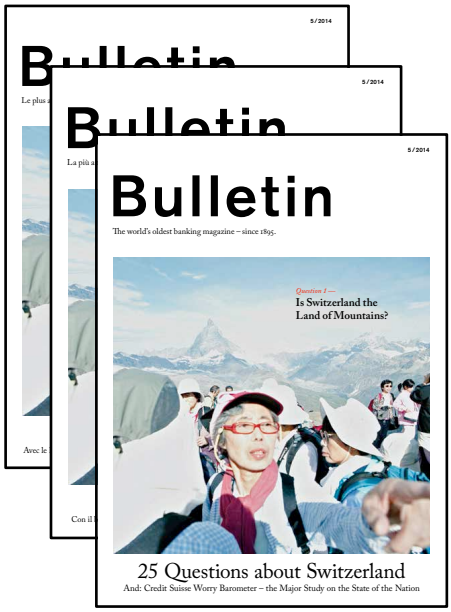
Boucle panoramique à deux pas de Leysin

Les touristes de passage à Leysin n'ont d'yeux que pour ses voisines d'Aï et de Mayen. La Tour de Famelon est pour les randonneurs à skis le clou d'une jolie course en boucle.
De Stéphane Maire



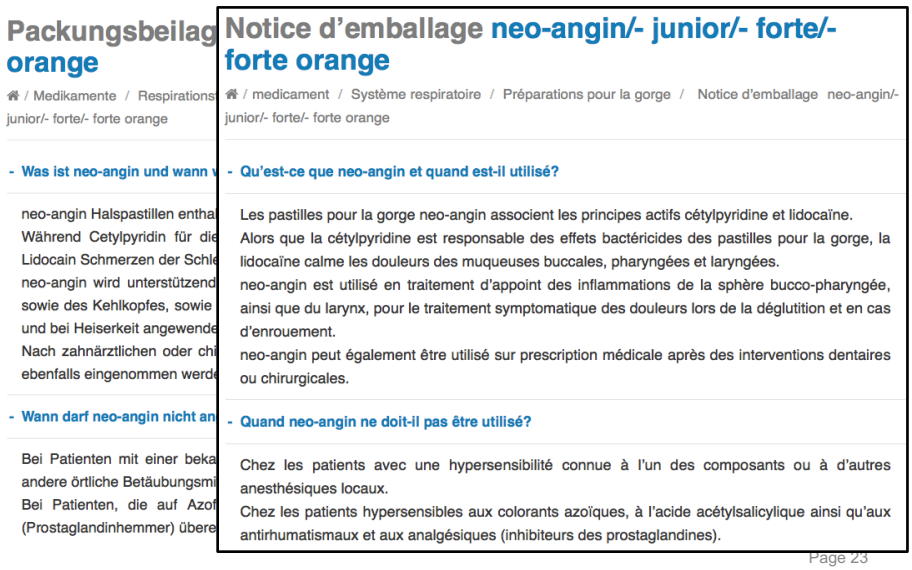
Vers l'W, la vue porte jusqu'aux géants de l'Oberland bernois: Eiger, Mönch, Jungfrau, Blüemlisalp et Doldenhorn (de g. à d.). Au premier plan, le Mont d'Or. © Stéphane Maire

Gegen Westen reicht die Aussicht bis zu den Riesen der Berner Alpen: Eiger, Mönch, Jungfrau, Blüemlisalp und Doldenhorn (von links nach rechts). Im Vordergrund der Mont d'Or. © Stéphane Maire



Medi-Notice Corpus

Medi-Notice Corpus



Packungsbeilag
orange

/ Medikamente / Respirations
junior/- forte/- forte orange

- Was ist neo-angin und wann v

neo-angin Halspastillen entha
Während Cetylpyridin für die
Lidocain Schmerzen der Schle
neo-angin wird unterstützend
sowie des Kehlkopfes, sowie
und bei Heiserkeit angewende
Nach zahnärztlichen oder chi
ebenfalls eingenommen werde

- Wann darf neo-angin nicht an

Bei Patienten mit einer beka
andere örtliche Betäubungsmi
Bei Patienten, die auf Azof
(Prostaglandinhemmer) übere

Notice d'emballage neo-angin/- junior/- forte/- forte orange

/ medicament / Système respiratoire / Préparations pour la gorge / Notice d'emballage neo-angin/- junior/- forte/- forte orange

- Qu'est-ce que neo-angin et quand est-il utilisé?

Les pastilles pour la gorge neo-angin associent les principes actifs cétylpyridine et lidocaïne.
Alors que la cétylpyridine est responsable des effets bactéricides des pastilles pour la gorge, la lidocaïne calme les douleurs des muqueuses buccales, pharyngées et laryngées.
neo-angin est utilisé en traitement d'appoint des inflammations de la sphère bucco-pharyngée, ainsi que du larynx, pour le traitement symptomatique des douleurs lors de la déglutition et en cas d'enrouement.
neo-angin peut également être utilisé sur prescription médicale après des interventions dentaires ou chirurgicales.

- Quand neo-angin ne doit-il pas être utilisé?

Chez les patients avec une hypersensibilité connue à l'un des composants ou à d'autres anesthésiques locaux.
Chez les patients hypersensibles aux colorants azoïques, à l'acide acétylsalicylique ainsi qu'aux antirhumatismaux et aux analgésiques (inhibiteurs des prostaglandines).

Page 23

University of Zurich[UZH]

Department of Computational Linguistics

Medi-Notice Corpus DE-FR at UZH

- patient information (4360 leaflets)
  - DE: 6.4 million – FR: 7.6 million tokens
  - DE: 67'478 types – FR: 46'604 types

- medical subject information (4114 leaflets)
  - DE: 14.4 million – FR: 18.3 million tokens
  - DE: 186'308 types – FR: 98'487 types

by Andrea Fritz (MLTA Master Project, Dec. 2016)

University of Zurich[UZH]

Department of Computational Linguistics

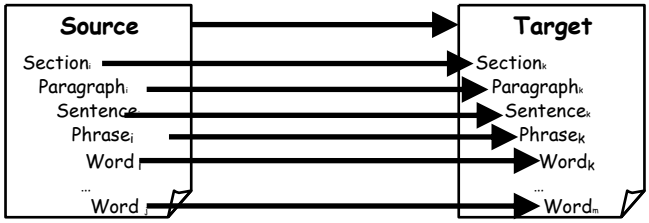
Text Alignment for Statistical MT

Parallel texts (= bitexts)

- Same content is available in several languages
- Official documents of countries with multiple official languages → literal, consistent

Alignment

- Paragraph to paragraph, sentence to sentence, word to word



University of Zurich[UZH]

Department of Computational Linguistics

1. Document Alignment

Given a document collection

- e.g. the yearbooks of the SAC

determine which document D1 in language L1 is a translation of document D2 in language L2.

Note: the document collection can be the web.

University of Zurich[UZH]

Department of Computational Linguistics

Document Alignment

Method:

- Compare the author names
- Compare the lengths of the documents (in number of characters)
- Compare the document structure (in number of paragraphs and sentences)
- Compare anchor words (e.g. numbers, names)

University of Zurich[UZH]

Department of Computational Linguistics

Example: DIE ALPEN 1965 DE - FR

Himalaya-Chronik 1963	G.O. Dyhrenfurth	1 DE	
La Pointe Courbe	Vériene Mettler	1 FR	Gemälde
Eine Besteigung des Fujiyama	W. K. Rieben	9 DE	
Der Stetind - das Matterhorn des Nordens	Henrik Bierberg	12 DE	
Die Bewässerungsanlagen von Ausserberg im Wallis	Ernst Lautenschlager	16 DE	
Eine Viertelstunde tödlicher Gewissheit	Hermann Roth	24 DE	
Das unverstandene Bergsteigen	Karl Greitbauer	26 DE	

Chronique himalayenne 1963	G.-O. Dyhrenfurth	1 FR	
La Pointe Courbe (Verbier)	Vérène Mettler	1 FR	peinture à l'huile
Eine Besteigung des Fujiyama	W. K. Rieben	9 DE	
Le Stetind - Le Cervin de la Norvège	Henrik Bierberg	12 FR	
Le système d'irrigation d'Ausserberg en Valais	Ernst Lautenschlager	16 FR	
Un quart d'heure sous le souffle de la mort	Hermann Roth	23 FR	
Das unverstandene Bergsteigen	Karl Greitbauer	26 DE	

University of Zurich[UZH]

Department of Computational Linguistics

2. Sentence Alignment

Given a pair of translated documents

– e.g. a pair of articles in the Text+Berg corpus

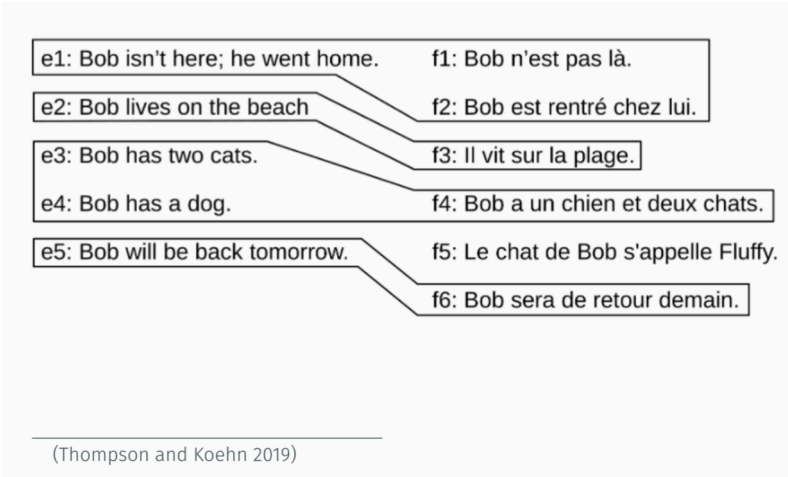
determine which sentence S1 in document D1 is a translation of which sentence S2 in document D2.

– Many-to-one sentence alignments are possible.
– Zero alignments are possible.
– We assume that the sentences in both documents are **in the same order**.

University of Zurich[UZH]

Department of Computational Linguistics

Toy Example



(Thompson and Koehn 2019)

1:n alignments

	English	German	Spanish
1	I hear MEPs who, I think, still believe in the effectiveness, honour and values of Europe, as well as feeling a certain pride in being European.	Europaabgeordnete, die meiner Meinung nach doch Grundsätze wie Effizienz und Ehre sowie die Wertvorstellungen Europas hochhalten und einen gewissen Stolz empfinden, Europäer zu sein – diese Abgeordneten höre ich ständig lamentieren und ein Sündenbekenntnis ablegen, dass an alledem im Grunde Europa schuld sei.	He escuchado las intervenciones de diputados al PE que, desde mi punto de vista, aún creen en la eficacia, el honor y los valores de Europa y que además sienten cierto orgullo de ser europeos.
2	I hear them constantly complaining and apologising.		Les he oído quejarse y pedir disculpas de un modo constante.
3	Basically this is all meant to be Europe's fault.		Todo esto significa esencialmente que es culpa de Europa y no puedo aceptarlo.
4	I do not accept that.	Dem stimme ich nicht zu.	

Die Bewässerungsanlagen von Ausserberg im Wallis		Le système d' irrigation d' Ausserberg en Valais
VON ERNST LAUTENSCHLAGER , BASEL	→	PAR ERNST LAUTENSCHLAGER , BALE
Mit 3 Bildern (25-27) und 1Kartenskizze Das Gelände der Gemeinde Ausserberg umfasst die Berglehnen zwischen dem Bietsch- und dem Baltschiedertal an der Südrampe der Lötschbergbahn .	→ ↘	avec 3 illustrations (25-27) et une carte-esquisse .
Im Mittelalter wurde dieses Gebiet Bischofsberg genannt .		Le territoire de la commune d' Ausserberg comprend les pentes situées entre les vallées de Bietsch et de Baltschieder à la rampe sud du chemin de fer du Lötschberg .
Seine fünf selbständigen Gemeinden führten ein gemeinsames Wappen :		Cette région s' appelait Bischofsberg au Moyen Age .
ein Kreuz auf grünem Grund , umgeben von 5 Sternen , überragt vom Doppeladler .		Ses cinq communes indépendantes portaient les mêmes armoiries :
Im 17. Jahrhundert verloren die fünf alten Gemeinden ihre Selbständigkeit .		une croix sur champ de sinopie , entourée de cinq étoiles , surmontée de l' aigle bicéphale .
Als neues Gemeinwesen für das gesamte Gebiet trat Ausserberg , « Mons Exterior » , auf , unter Beibehaltung des alten Wappens .		Au XVIIe siècle , les cinq communes perdirent leur indépendance .
		La nouvelle commune se nomma Ausserberg « Mons exterior » et conserva les anciennes armoiries .

Die Bewässerungsanlagen von Ausserberg im Wallis	→	Le système d' irrigation d' Ausserberg en Valais
VON ERNST LAUTENSCHLAGER , BASEL	→	PAR ERNST LAUTENSCHLAGER , BALE
Mit 3 Bildern (25-27) und 1Kartenskizze Das Gelände der Gemeinde Ausserberg umfasst die Berglehnen zwischen dem Bietsch- und dem Baltschiedertal an der Südrampe der Lötschbergbahn .	→	avec 3 illustrations (25-27) et une carte-esquisse . Le territoire de la commune d' Ausserberg comprend les pentes situées entre les vallées de Bietsch et de Baltschieder à la rampe sud du chemin de fer du Lötschberg .
Im Mittelalter wurde dieses Gebiet Bischofsberg genannt .	→	Cette région s' appelait Bischofsberg au Moyen Age .
Seine fünf selbständigen Gemeinden führten ein gemeinsames Wappen :	→	Ses cinq communes indépendantes portaient les mêmes armoiries :
ein Kreuz auf grünem Grund , umgeben von 5 Sternen , überragt vom Doppeladler .	→	une croix sur champ de sinopie , entourée de cinq étoiles , surmontée de l' aigle bicéphale .
Im 17. Jahrhundert verloren die fünf alten Gemeinden ihre Selbständigkeit .	→	Au XVIIe siècle , les cinq communes perdirent leur indépendance .
Als neues Gemeinwesen für das gesamte Gebiet trat Ausserberg , « Mons Exterior » , auf , unter Beibehaltung des alten Wappens .	→	La nouvelle commune se nomma Ausserberg « Mons exterior » et conserva les anciennes armoiries .

University of Zurich[UZH]

Department of Computational Linguistics

Sentence Alignment

can be based on length comparison (# of characters).

can be based on anchor words:

- numbers
  - but sometimes Arabic vs. Roman numbers (*17* vs. *XVII*)
- names
  - but sometimes translated (*Matterhorn* vs. *Cervin*)
- cognates
  - DE: *Lokomotive* – FR: *locomotive*

Important assumption: the sentences in both texts are in the same order!

University of Zurich[UZH]

Department of Computational Linguistics

Example Tool for Sentence Alignment

InterText / HunAligner

freely available at

http://wanthalf.saga.cz/intertext

Reference

– Vondřička, Pavel (2014): "Aligning parallel texts with InterText" In: *Proceedings of the Conference on Language Resources and Evaluation (LREC'14)*. p. 1875-1879.

University of Zurich[UZH]

Department of Computational Linguistics

HunAlign – Sentence Aligner

aligns bilingual text on the sentence level.

the input is tokenized and sentence-segmented text in two languages.

does not deal with changes of sentence order (= no crossing alignments).

is written in C++ (around 2005)

http://mokk.bme.hu/en/resources/hunalign/

University of Zurich[UZH]

Department of Computational Linguistics

HunAlign – Sentence Aligner

With a bilingual dictionary, HunAlign

– uses the dictionary together with Gale-Church (1991) sentence-length information.

Without a dictionary, HunAlign

1. uses sentence-length information,
2. builds an automatic dictionary based on this alignment,
3. realigns the text in a second pass, using the automatic dictionary.

University of Zurich[UZH]

Department of Computational Linguistics

HunAlign realign option

If this option is set, the alignment is built in 3 phases.

After an initial alignment, the algorithm heuristically adds items to the dictionary based on co-occurrences in the parallel sentences (bi-sentences).

Then it re-runs the alignment process based on this larger dictionary.

This option is recommended to achieve the highest possible alignment quality.

This option roughly triples the running time while the improvements are typically small.

University of Zurich[UZH]

Department of Computational Linguistics

HunAlign – Handling XML

If a line consists of a single <p> token, it is treated as a paragraph delimiter.

Paragraph separators are treated as virtual sentences, the aligner tries to match these with each other and never aligns them with a real sentence.

University of Zurich[UZH]

Department of Computational Linguistics

InterText Features

- import and export of ready-made alignments
- creating new alignments from plain text or XML files (sentence-segmented or not)
  - possibility to (re)align any part of the alignment with the automatic aligner
- full editing possibilities of the alignment and the element contents
- search & replace functionality (incl. regular expressions)

University of Zurich[UZH]

Department of Computational Linguistics

InterText / HunAligner

credit_suisse_news_corpus_en_it – InterText

	M	en	it
1		EU enlargement : now the people must be heard	Allargamento dell' UE : la parola al popolo
2		17.03.2003	17.03.2003
3		The first decision has been made . The Maltese voted for entry into the EU on March 9 , yet this is not yet signed and sealed , as opponents to EU membership in parliament could still turn the tables .	La prima decisione è stata presa : il 9 marzo i cittadini maltesi si sono espressi a favore dell' adesione all' UE . Questa decisione non è però stata suggellata , poiché in Parlamento i fautori del « no » potrebbero ancora rovesciare la situazione .
4		Meanwhile , tension is mounting in the other candidate countries over the forthcoming EU referenda .	Nel frattempo anche negli altri paesi candidati sale la tensione prima delle consultazioni popolari sull' adesione all' UE .
5		The Maltese vote for entry into the European Union was carried by a small majority .	Una stretta maggioranza della popolazione dell' isola di Malta si è pronunciata positivamente in merito all' adesione all' Unione europea .
6		In a referendum on March 9 , 53.6 percent voted for entry into the EU , with 46.4 percent against .	Nel referendum del 9 marzo il 53,6 per cento degli elettori ha votato per l' ingresso nell' UE , il 46,4 per cento contro .
7		91 percent of the electorate voted .	La partecipazione al voto si è attestata al 91 per cento .
8		This was a vote of confidence in the Maltese Prime Minister Fenech Adami , who supports entry into the EU .	Con questo risultato la popolazione maltese ha sostenuto il primo ministro Fenech Adami che si impegna a favore dell' adesione .
9			La decisione definitiva spetta al Parlamento
10		Mr Adami announced the date for a general election immediately after the referendum victory - April 12 , 2003 , just four days before the EU candidate countries are due to sign the accession treaty with the EU .	Dopo la vittoria alle urne , Adami ha immediatamente indetto nuove elezioni per il Parlamento . Esse avranno luogo il 12 aprile 2003 , prima della firma del Trattato d' adesione con l' UE del 16 aprile da parte dei paesi candidati .

41

University of Zurich[UZH]

Department of Computational Linguistics

Formats for Parallel Corpora

= e.g. Intertext output options:

parallel files with line-by-line alignment

TMX = Translation Memory eXchange = XML format for sentence-aligned text

TMX Example

```
<tmx version="1.4b">
    <header creationtool="InterText" creationtoolversion="1.0" datatype="PlainText"
segtype="block" adminlang="en-us" srclang="en" o-tmf="XML aligned text"></header>
<body>
<tu><prop type="x-sentbreak">|#|</prop>
    <tuv xml:lang="en"><seg>"Switzerland needs entrepreneurs.</seg></tuv>
    <tuv xml:lang="de"><seg>"Die Schweiz braucht Unternehmer.</seg></tuv>
</tu>
<tu><prop type="x-sentbreak">|#|</prop>
    <tuv xml:lang="en"><seg>It's where the future lies."</seg></tuv>
    <tuv xml:lang="de"><seg>Darin liegt die Zukunft."</seg></tuv>
</tu>
<tu><prop type="x-sentbreak">|#|</prop>
    <tuv xml:lang="en"><seg>Such, in a nutshell, was the credo of William A. de Vigier
(1912-2003), who established the foundation of the same name that awards up to five highly
valuable advancement awards of 100,000 Swiss francs each to young Swiss
entrepreneurs.</seg></tuv>
    <tuv xml:lang="de"><seg>So kurz und klar war das Credo von William A. de Vigier
(1912-2003), dem Gründer der gleichnamigen Stiftung, die jedes Jahr bis zu fünf hoch
dotierte Förderpreise (je 100 000 Franken) an Schweizer Jungunternehmer
vergibt.</seg></tuv>
</tu>
```


University of Zurich[UZH]

Department of Computational Linguistics

XLIFF

= XML Localization Interchange File Format

- a common format for CAT tool exchange
- better than TMX ???
- advantage: all translations are stored in one file.
- Example: if a company translates English source text into 5 different languages, all of those translations would be stored in **one** XLIFF file.

University of Zurich[UZH]

Department of Computational Linguistics

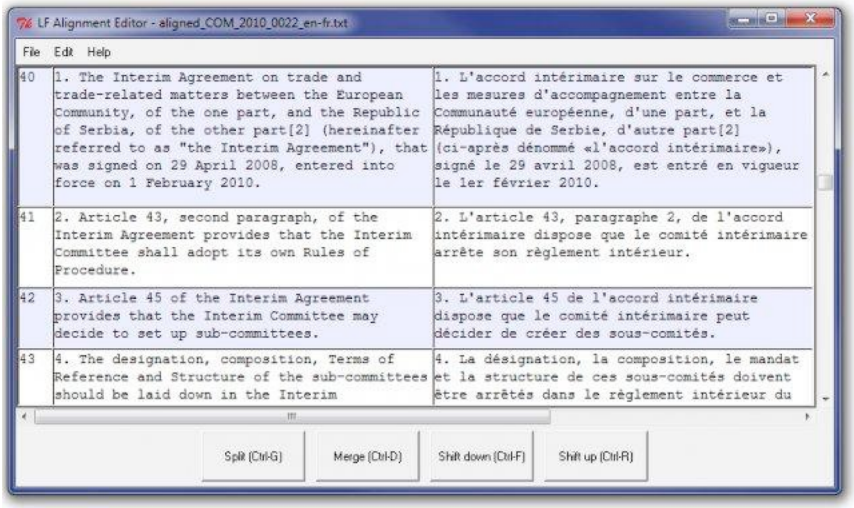
LF Aligner

- recommended by Martin Kappus
- “LF Aligner helps translators create translation memories from texts and their translations. It relies on **Hunalign** for automatic sentence pairing.
- Input: txt, doc, docx, rtf, pdf, html.
- Output: tab delimited txt, **TMX** and xls.”
- https://sourceforge.net/projects/aligner/

University of Zurich[UZH]

Department of Computational Linguistics

LF Aligner



University of Zurich[UZH]

Department of Computational Linguistics

Sentence Alignment (Alternative Method)

can be based on Machine Translation

Idea:

- Machine Translation of S1_L1 → S_{MT}_L2
- If S_{MT}_L2 is "close to" S2_L2,
- then S1_L1 and S2_L2 will be aligned.

This method is well suited for "noisy" text.

University of Zurich[UZH]

Department of Computational Linguistics

BLEU-Align (by Rico Sennrich)

Sentence Alignment via MT Example:

1. Text in English
2. Text in German → MT to English
3. Comparison of the *English_orig* sentences to the *English_MT* sentences. If the sentence pair has a high BLEU score, then keep it as anchors.
4. Continue the comparison of the *English_orig* sentences to the *English_MT* sentences. Accept also neighboring sentence pairs with medium BLEU scores.
5. Option: Reverse the MT direction and compare again.

University of Zurich[UZH]

Department of Computational Linguistics

3. Word Alignment

Given a pair of translated sentences

– determine which word(s) w1 in sentence S1 is/are a translation of which word(s) w2 in sentence S2.
  – Many-to-one word alignments are possible.
  – We cannot assume that the words in both sentences are in the same order.

University of Zurich[UZH]

Department of Computational Linguistics

Word Alignment Example

Im	→	Au
17.	→	XVIIe
Jahrhundert	→	siècle
verloren		,
die	→	les
fünf	→	cinq
alten		communes
Gemeinden		perdirent
ihre	→	leur
Selbständigkeit	→	indépendance
.	→	.

Statistical Word Alignment

Statistical Word Alignment is based on co-occurrence counts.

Statistical Phrase Alignment (= Word Sequence Alignment) is based on Word Alignment.

University of Zurich[UZH]
Department of Computational Linguistics

Word Alignment

DE	FR
Gegen Abend hielten wir dann Einzug in das komfortable **Haus** der Berliner.	Vers le soir nous entrons dans la confortable maison des Berlinois.
Eine seiner alten Tanten, die immer gerne das ganze **Haus** regierte, nahm ihn leidenschaftlich ins Gebet:	Une de ses vieilles tantes, qui régentait volontiers toute la maison, le prit vivement à partie:
Als Alpinist und Tourist weiss man ja ein **Haus** in einsamer Gegend sehr zu schätzen.	En tant qu'alpiniste ou touriste, on sait apprécier la présence d'un bâtiment dans une région inhabitée.
Das **Haus** war unverschlossen, aber kein Mensch kam auf unser Lärmen herbei.	La maison n'était pas fermée, mais personne n'apparut à nos appels.

University of Zurich[UZH]

Department of Computational Linguistics

Word Alignment

DE	FR
Gegen Abend hielten wir dann Einzug in das komfortable **Haus** der Berliner.	Vers le soir nous entrons dans la confortable **maison** des Berlinois.
Eine seiner alten Tanten, die immer gerne das ganze **Haus** regierte, nahm ihn leidenschaftlich ins Gebet:	Une de ses vieilles tantes, qui régentait volontiers toute la **maison**, le prit vivement à partie:
Als Alpinist und Tourist weiss man ja ein **Haus** in einsamer Gegend sehr zu schätzen.	En tant qu'alpiniste ou touriste, on sait apprécier la présence d'un **bâtiment** dans une région inhabitée.
Das **Haus** war unverschlossen, aber kein Mensch kam auf unser Lärmen herbei.	La **maison** n'était pas fermée, mais personne n'apparut à nos appels.

University of Zurich[UZH]

Department of Computational Linguistics

textshuttle.ai

Thank you!

volk@cl.uzh.ch