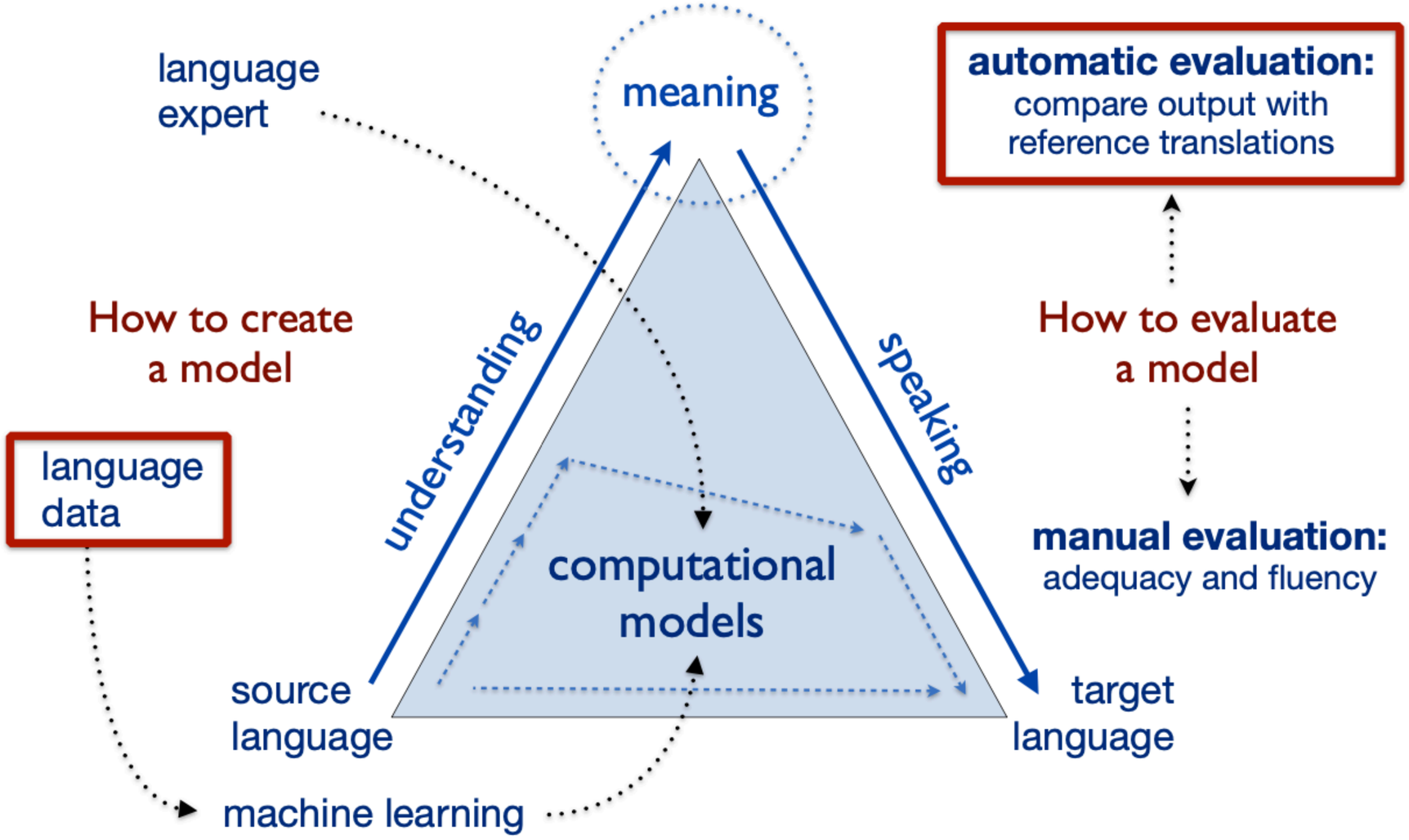
Machine Translation

Session 2: Training data for NMT

Sharid Loáiciga — May 6th, 2020

Slides credits: Jörg Tiedemann
Figures from https://www.morganclaypool.com/doi/abs/10.2200/S00367ED1V01Y201106HLT014

MT in a nutshell



The advantage of data driven MT

Human Translations Naturally Appear

- no need for artificial annotation
- can be provided by non-experts

Implicit Linguistics

- learnable parameters instead of fixed formalisms
- translation knowledge is in the data

Constant Learning is Possible

- feed with new data as they appear
- quickly adapt to new domains and language pairs

Data-driven MT

learn the
unknown code

When I look at an article in Russian, I say:
This is really written in English,
but it has been coded in some strange symbols.
I will now proceed to decode.

[Weaver, 1947, 1949]

Data-driven MT

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

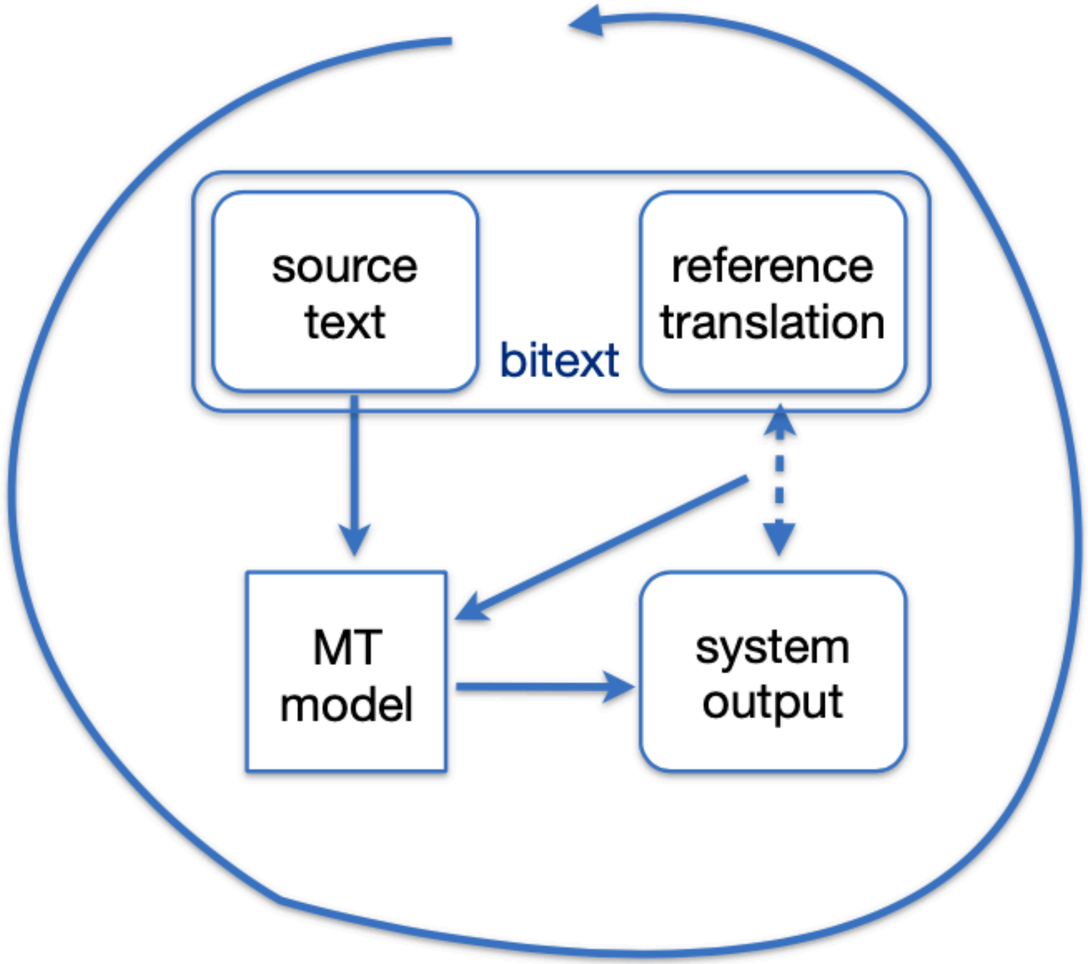
Probabilistic Model: P(**e**|**c**)

Search problem (decoding):

$$\boldsymbol{e}^* = \underset{\boldsymbol{e}}{argmax}\ P(\boldsymbol{e}|\boldsymbol{c})$$

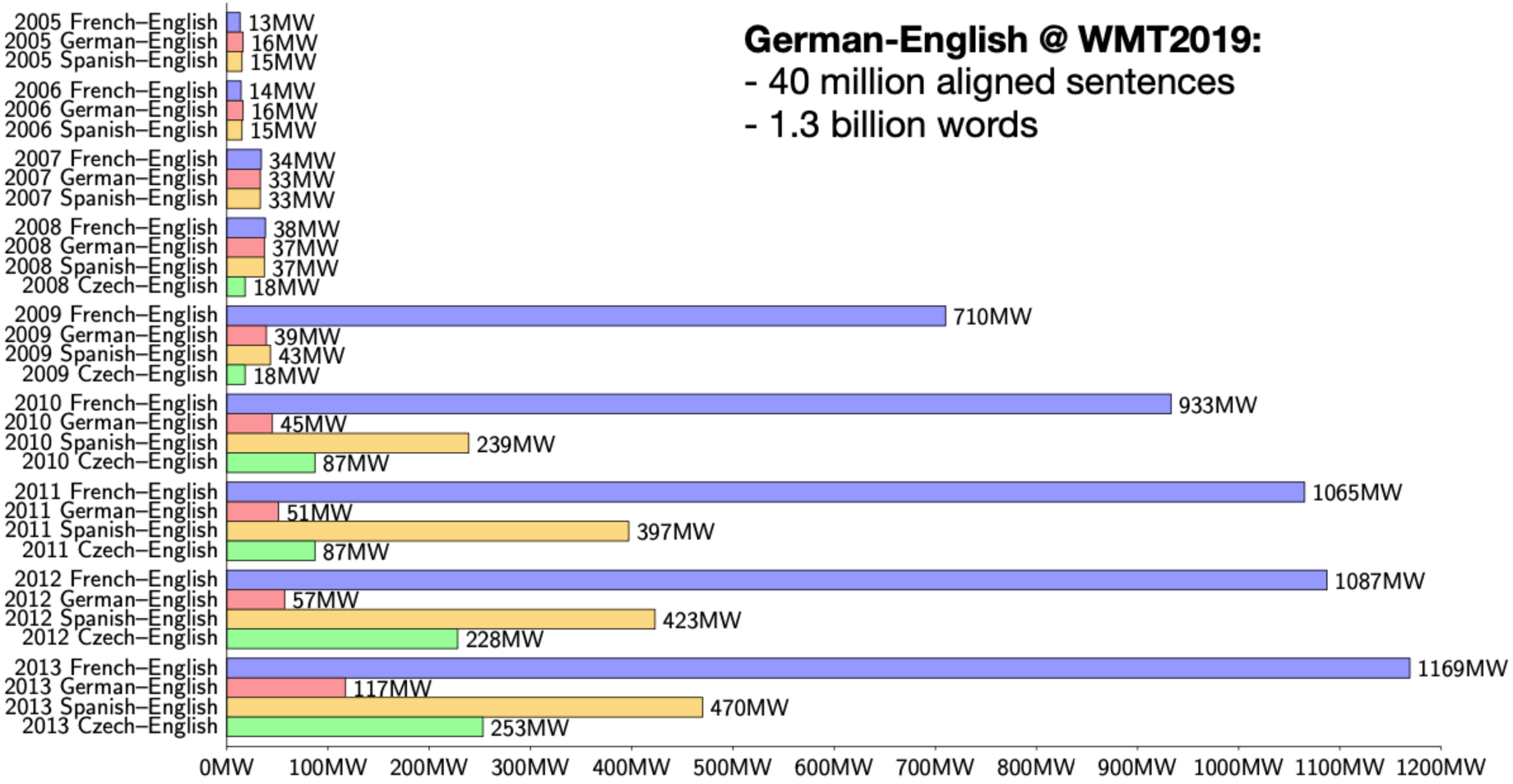
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Supervised training in MT



repeat until satisfied

Growing data sets at WMT



(Source: Philipp Koehn)

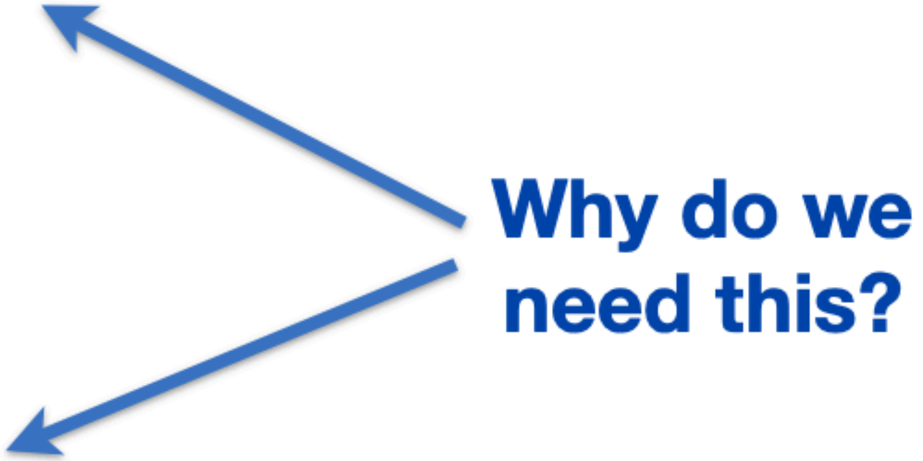
How to create training data

Collecting and Pre-Processing

- web-crawling, ...
- converting
- detect sentence and word boundaries, lemmatisation ...

Aligning

- document alignment
- paragraph (optional)
- sentence alignment



Why do we need pre-processing

Remember the probabilistic model?

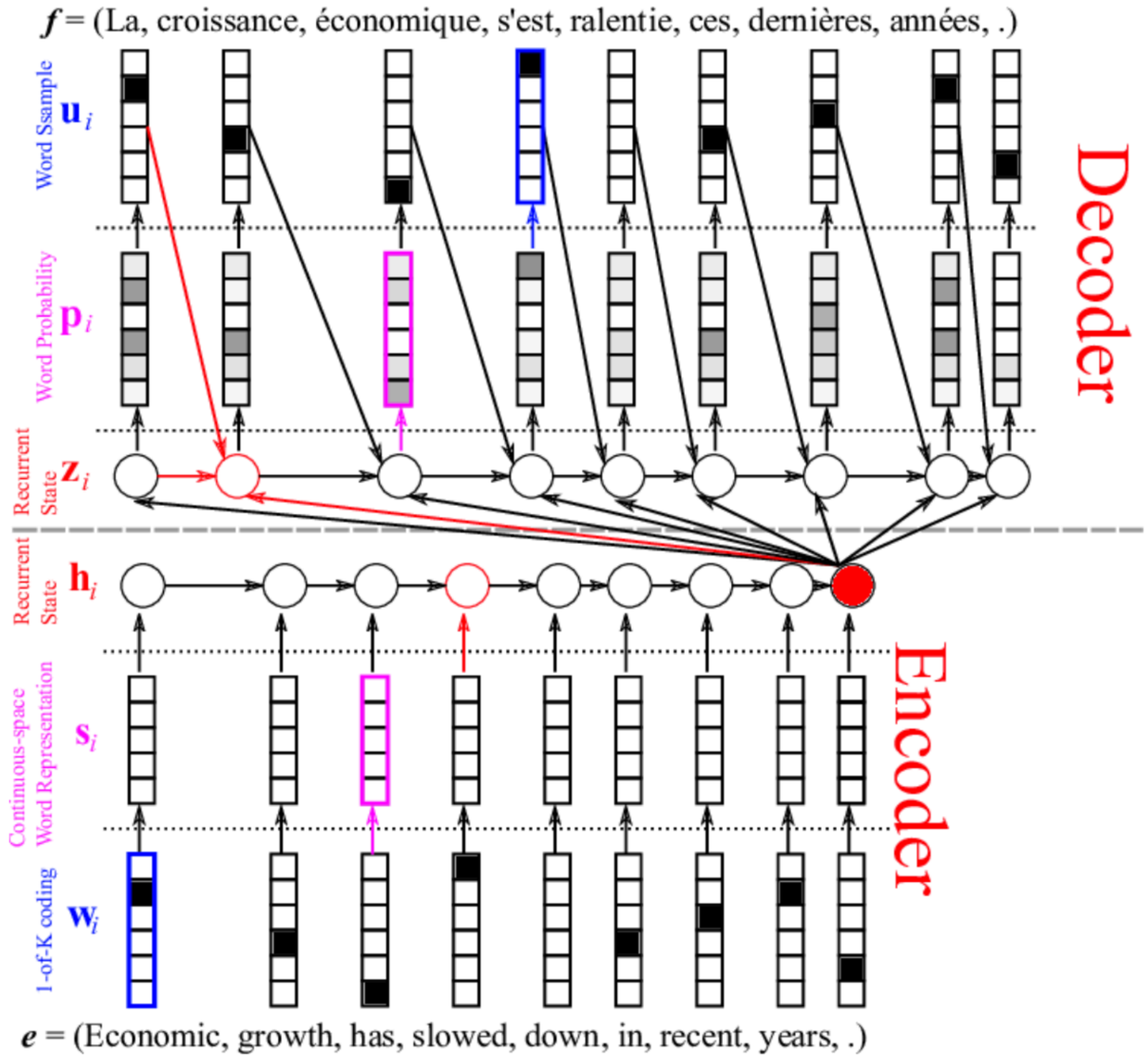
Translation = decoding

$$\mathbf{e}^* = \underset{\mathbf{e}}{argmax}\ P(\mathbf{e}|\mathbf{c})$$

Running over all possible English texts to find the best translation given our Chinese text!

How do we know the probability that an English text is a good translation of a Chinese one?

Decomposition in NMT



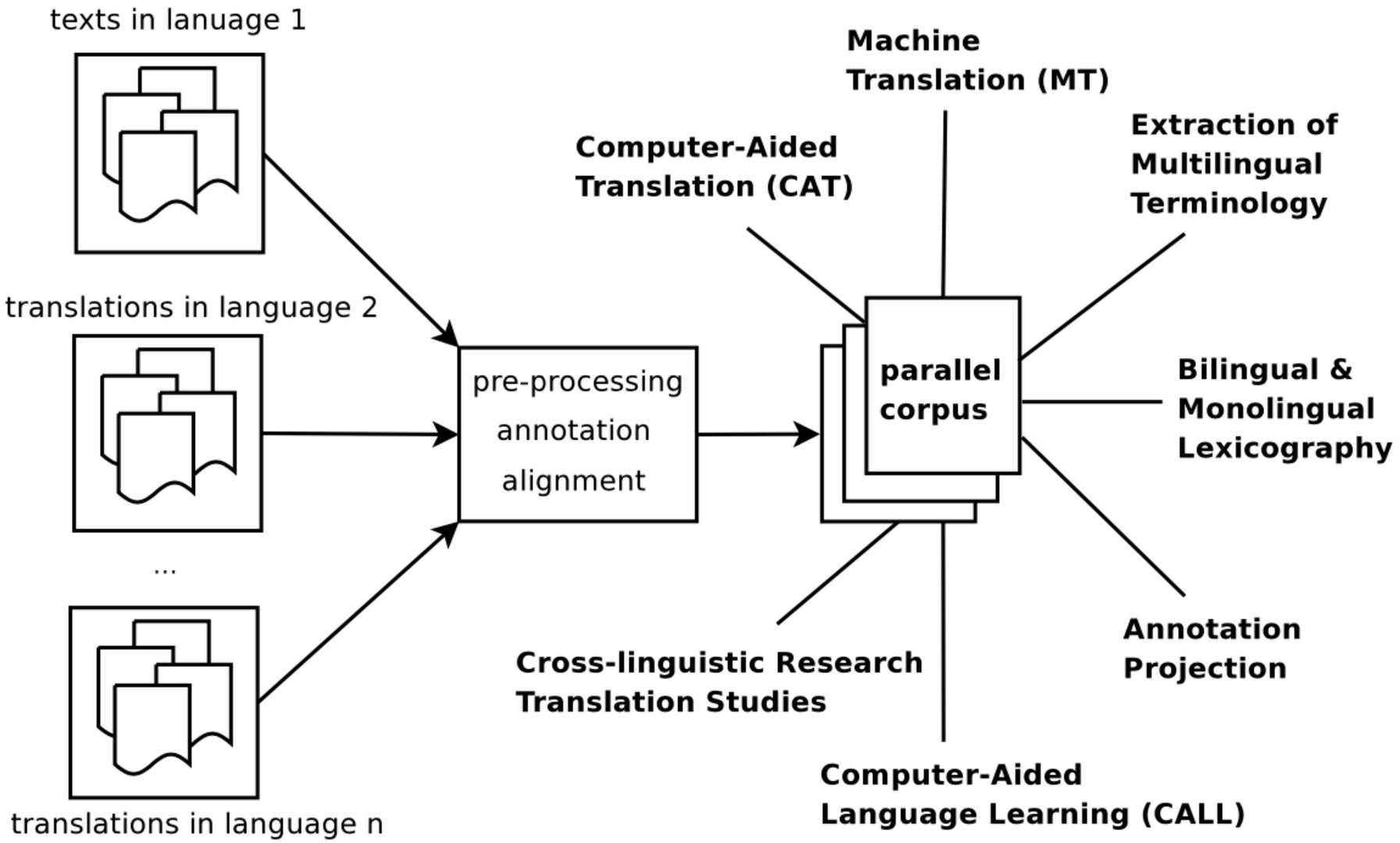
https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-2/

Sentence-aligned parallel corpora

Essential training data = tokenized and aligned bitexts

what is more , the relevant cost dynamic is completely under control.	im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle .
sooner or later we will have to be sufficiently progressive in terms of own resources as a basis for this fair tax system .	früher oder später müssen wir die notwendige progressivität der eigenmittel als grundlage dieses gerechten steuersystems zur sprache bringen .
we plan to submit the first accession partnership in the autumn of this year .	wir planen , die erste beitrittspartnerschaft im herbst dieses jahres vorzulegen .
it is a question of equality and solidarity .	hier geht es um gleichberechtigung und solidarität .
the recommendation for the year 1999 has been formulated at a time of favourable developments and optimistic prospects for the european economy .	die empfehlung für das jahr 1999 wurde vor dem hintergrund günstiger entwicklungen und einer für den kurs der europäischen wirtschaft positiven perspektive abgegeben .
that does not , however , detract from the deep appreciation which we have for this report .	im übrigen tut das unserer hohen wertschätzung für den vorliegenden bericht keinen abbruch .

Building and using parallel corpora



Collecting parallel corpora

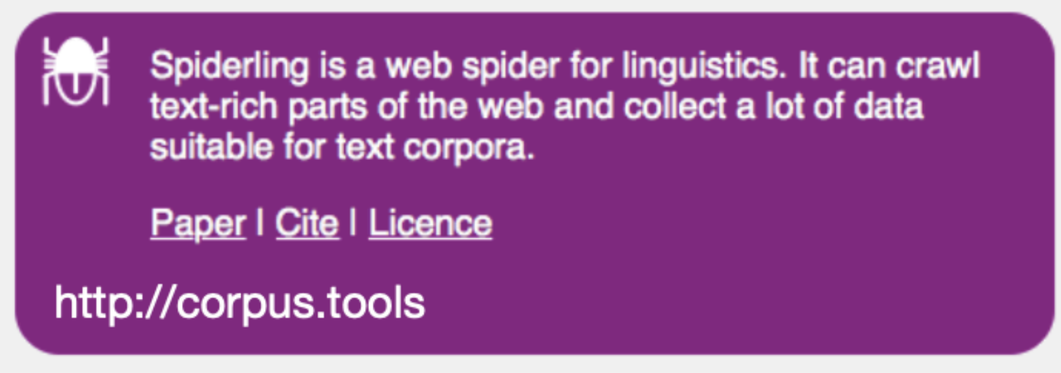
Where does the data come from?

- multilingual legislation
- public sector in multilingual communities
- international entertainment (movies, books, …)
- global products (localisation), translated websites
- translation agencies (you wish …)

What are the problems?

- copyright, GDPR issues
- incompleteness, noise, language coverage

Collecting parallel corpora



ParaCrawl

langid.py

Google Chrome CLD

GNU Wget

https://scrapy.org

Bitextor

Scrapy

OPUS

http://www.statmt.org/europarl/

http://www.statmt.org/wmt19/

http://opus.nlpl.eu

Converting data

https://tika.apache.org

http://corpus.tools

JusText is a HTML boilerplate removal tool. It can strip navigation links, headers, footers, etc. from HTML pages and leave just regular text containing full sentences.

Paper | Cite | Licence

Chared is a tool for detecting the character encoding of a text in a known language. It contains models for a wide range of languages.

Paper | Cite | Licence

http://www.xpdfreader.com

Onion (ONe Instance ONly) is a de-duplicator for large collections of texts. It can measure the similarity of paragraphs or whole documents and drop duplicate ones based on the threshold you set.

Paper | Cite | Licence

https://poppler.freedesktop.org

Sentences and tokenization

Rule-based approaches

- regular expressions and word lists
- language-specific rules
- moses-scripts

Data-driven methods

- trained on tokenized texts
- OpenNLP: https://opennlp.apache.org
- UDpipe: http://ufal.mff.cuni.cz/udpipe

Other pre-processing tools

- Text segmentation and morphological analysis
    - Japanese: MeCab, ChaSen
    - Chinese: http://nlpprogress.com/chinese/chinese_word_segmentation.html
    - UD languages: parsing, for example https://stanfordnlp.github.io/stanza/
- Text normalization (part of the Moses package)
    - Truecasing
    - Normalization

Sentence alignment

Task:

- align corresponding sentences to each other
  (may be sequences of sentences)

Assumption:

- sentence alignment can be done **monotonically** (no crossing links)

Challenges:

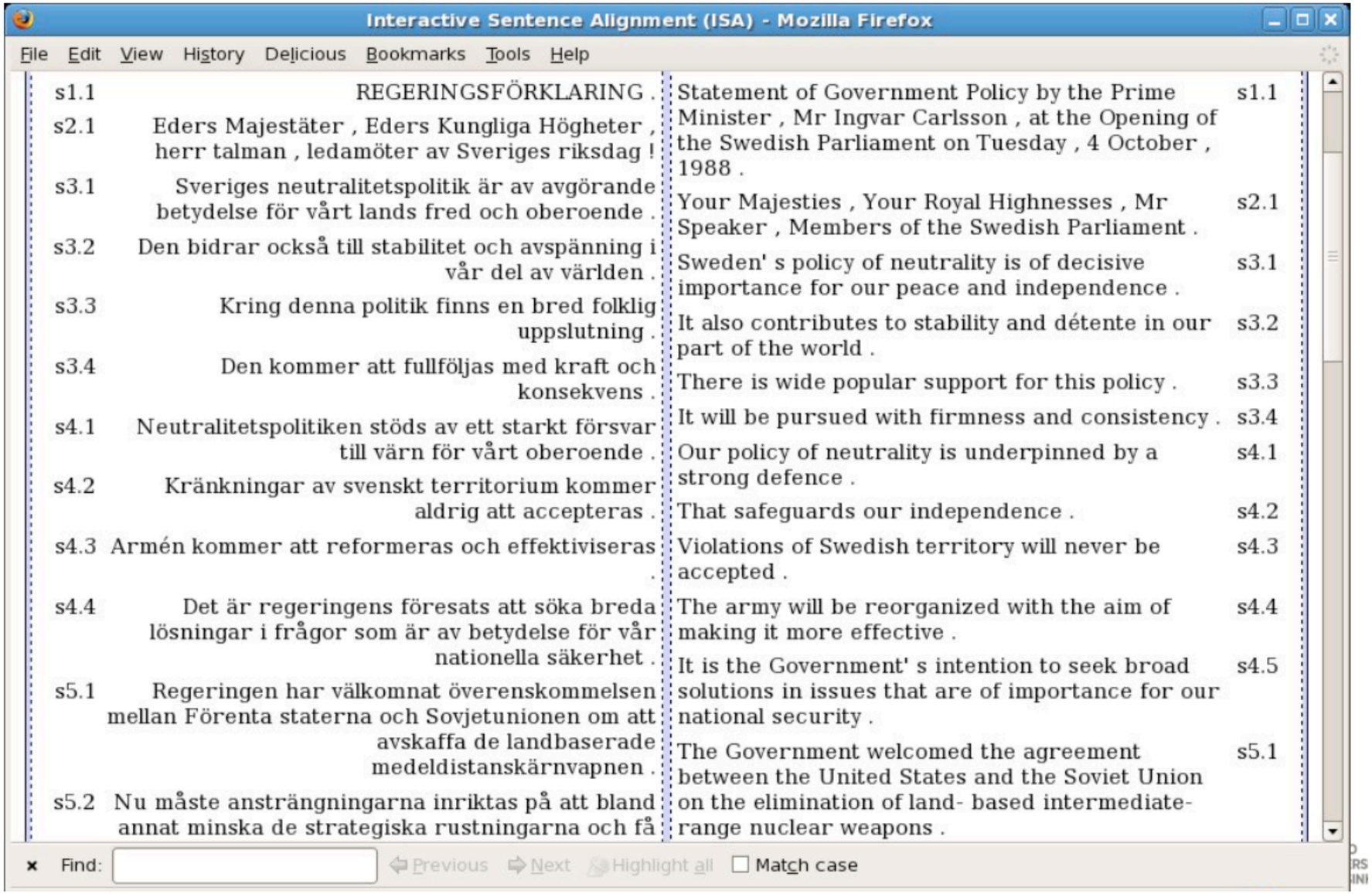
- non-1:1 alignments, insertions, deletions, incomplete translations

Many different ways to align sentences:

- (s1 → t1) (s2 → t2) (s3 → t3) ...
- (s1 → t1,t2) (s2 → t3) (s3 → 0) ...
- (s1,s2 → t1) (s3 → t2,t3) ...

  ...

Simple example



Simple example sentence-aligned

s1.1	REGERINGSFÖRKLARING .	Statement of Government Policy by the Prime Minister , Mr Ingvar Carlsson , at the Opening of the Swedish Parliament on Tuesday , 4 October , 1988 .	s1.1
s2.1	Eders Majestäter , Eders Kungliga Högheter , herr talman , ledamöter av Sveriges riksdag !	Your Majesties , Your Royal Highnesses , Mr Speaker , Members of the Swedish Parliament .	s2.1
s3.1	Sveriges neutralitetspolitik är av avgörande betydelse för vårt lands fred och oberoende .	Sweden' s policy of neutrality is of decisive importance for our peace and independence .	s3.1
s3.2	Den bidrar också till stabilitet och avspänning i vår del av världen .	It also contributes to stability and détente in our part of the world .	s3.2
s3.3	Kring denna politik finns en bred folklig uppslutning .	There is wide popular support for this policy .	s3.3
s3.4	Den kommer att fullföljas med kraft och konsekvens .	It will be pursued with firmness and consistency .	s3.4
s4.1	Neutralitetspolitiken stöds av ett starkt försvar till värn för vårt oberoende .	Our policy of neutrality is underpinned by a strong defence .	s4.1
		That safeguards our independence .	s4.2
s4.2	Kränkningar av svenskt territorium kommer aldrig att accepteras .	Violations of Swedish territory will never be accepted .	s4.3
s4.3	Armén kommer att reformeras och effektiviseras .	The army will be reorganized with the aim of making it more effective .	s4.4
s4.4	Det är regeringens föresats att söka breda lösningar i frågor som är av betydelse för vår nationella säkerhet .	It is the Government' s intention to seek broad solutions in issues that are of importance for our national security .	s4.5
s5.1	Regeringen har välkomnat överenskommelsen mellan Förenta staterna och Sovjetunionen om att avskaffa de landbaserade medeldistanskärnvapnen .	The Government welcomed the agreement between the United States and the Soviet Union on the elimination of land- based intermediate- range nuclear weapons .	s5.1

Try to align the following sentences

source language	ID	target language
Fooi Tiadii , hseatenis aoe iscesnaohtmutis emt eis Lsoih , xes ücis aot iisioohicsudiio .	1	Wo wes qmåheei ew io iqoeino tun wes nis iäotzotmöt äo lsoh .
Fooi Qitutiadii , eoi eoi lämgui aotisit Löoohsiodiit eeiooseggui .	2	Fo qitu tun lun euu eöee nis äo iämguio ew soliu .
Xu len toi iis ?	3	Wesu lun eio ogsåo ?
Xoi wiscsiouiui toi todi ?	4	Win tqsie eio ?
Eoi Qsoituis tehuio aot , toi tio eoi Tusegi Huuuit .	5	Qsätuisoe cisäuueei euu eiu wes Haet cituseggoooh .
Bcis güs ximdii Tüoei ?	6	Gös womlio tzoe ?
Ximdiit Hicuu ieuuio xos hicsudiio , eett xos tu iuxet wiseoiou ieuuio ?	7	Oik , wo wottui teoooohio
Oioo , xos leoouio eoi Xeisiiou .	8	Eiuue wes ooui Haet wisl , aueo ekäwamiot .
Eet xes oodiu Huuuit Xisl , tuoeiso Uiagimio ... ueis liyisio .	9	Fmmis usummeun .
Voe aotisi Bagheci citueoe eesoo , güs eoi liomaoh easdi Huuu eio Eänuo za geohio .	10	Wo wes uwaohoe euu citihse io einuo .
Csaeis Uiunet !	11	Gös Haet gösmåuimti .
	12	Csueis Uiunet .

Re-arranging the table

ID	source language	target language	ID
1	Fooi Tiadii , hseatenis aoe iscesnaohtmutis emt eis Lsoih , xes ücis aot iisioohicsudiio .	Wo wes qmåheei ew io iqoeino tun wes nis iäotzotmöt äo lsoh .	1
2	Fooi Qitutiadii , eoi eoi lämgui aotisit Löoohsiodiit eeiooseggui .	Fo qitu tun lun euu eöee nis äo iämguio ew soliu .	2
3	Xu len toi iis ?	Wesu lun eio ogsåo ?	3
4	Xoi wiscsiouiui toi todi ?	Win tqsie eio ?	4
5	Eoi Qsoituis tehuio aot , toi tio eoi Tusegi Huuuit .	Qsätuisoe cisäuueei euu eiu wes Haet cituseggoooh .	5
6	Bcis güs ximdii Tüoei ?	Gös womlio tzoe ?	6
7	Ximdiit Hicuu ieuuio xos hicsudiio , eett xos tu iuxet wiseoiou ieuuio ?		
8	Oioo , xos leoouio eoi Xeisiiou .	Oik , wo wottui teoooohio	7
9	Eet xes oodiu Huuuit Xisl , tuoeiso Uiagimio ... ueis liyisio .	Eiuue wes ooui Haet wisl , aueo ekäwamiot .	8
		Fmmis usummeun .	9
10	Voe aotisi Bagheci citueoe eesoo , güs eoi liomaoh easdi Huuu eio Eänuo za geohio .	Wo wes uwaohoe euu citihse io einuo .	10
		Gös Haet gösmåuimti .	11
11	Csaeis Uiunet !	Csueis Uiunet .	12

Decoding the example

ID	source language	target language	ID
1	Eine Seuche , grausamer und erbarmungsloser als der Krieg , war über uns hereingebrochen .	Vi var plågade av en epidemi som var mer hänsynslös än krig .	1
2	Eine Pestseuche , die die Hälfte unseres Königreiches dahinraffte .	En pest som kom att döda mer än hälften av riket .	2
3	Wo kam sie her ?	Vart kom den ifrån ?	3
4	Wie verbreitete sie sich ?	Vem spred den ?	4
5	Die Priester sagten uns , sie sei die Strafe Gottes .	Prästerna berättade att det var Guds bestraffning .	5
6	Aber für welche Sünde ?	För vilken synd ?	6
7	Welches Gebot hatten wir gebrochen , dass wir so etwas verdient hatten ?		
8	Nein , wir kannten die Wahrheit .	Nej , vi visste sanningen	7
9	Das war nicht Gottes Werk , sondern Teufelei ... oder Hexerei .	Detta var inte Guds verk , utan djävulens .	8
		Eller trolldom .	9
10	Und unsere Aufgabe bestand darin , für die Heilung durch Gott den Dämon zu fangen .	Vi var tvungna att besegra en demon .	10
		För Guds förlåtelse .	11
11	Bruder Thomas !	Broder Thomas .	12

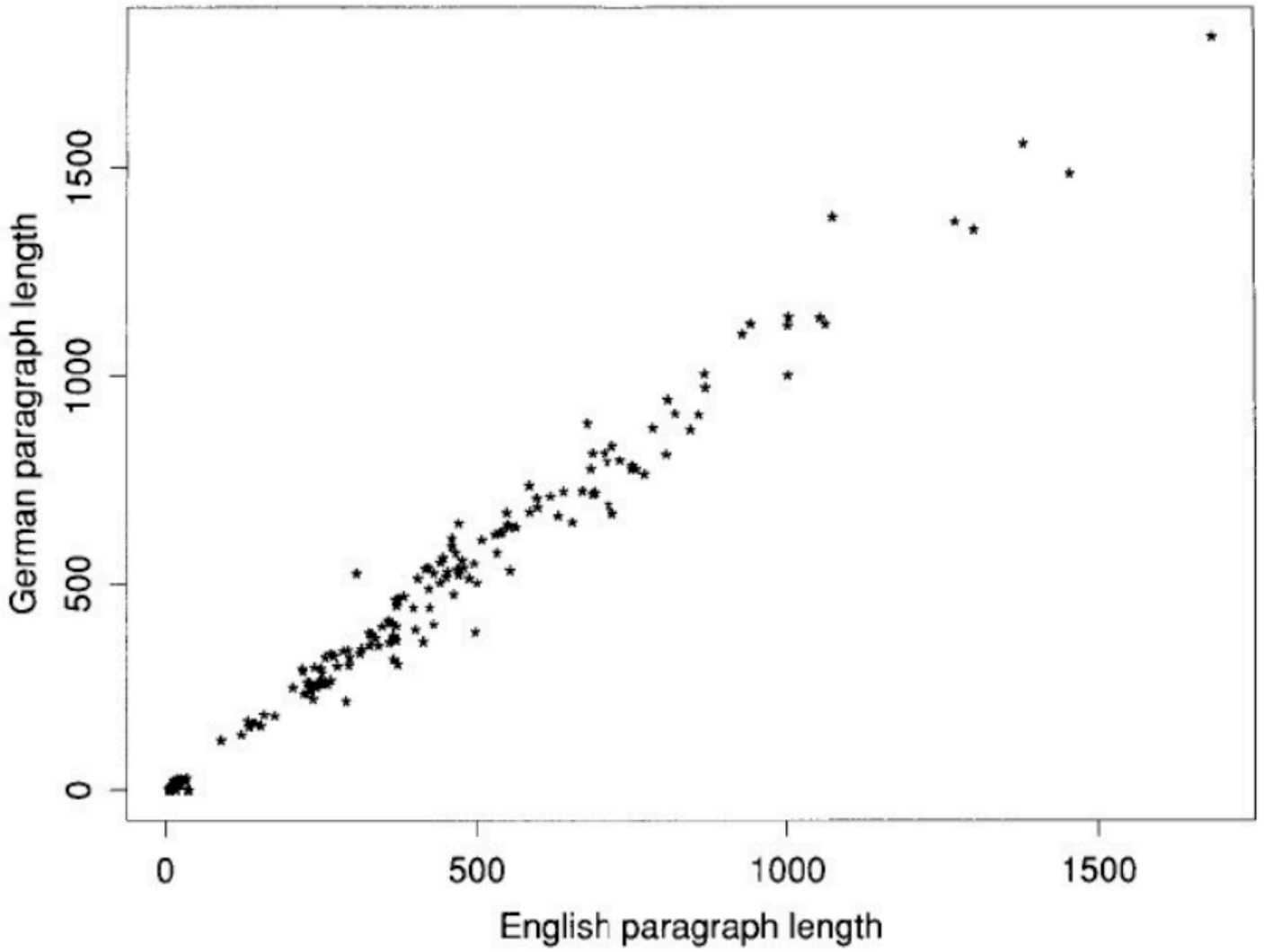
Automatic sentence alignment

Length-based methods: assumption = sentences (and sequences of sentences) that correspond to each other are also similar in length (characters or words) (more than others)

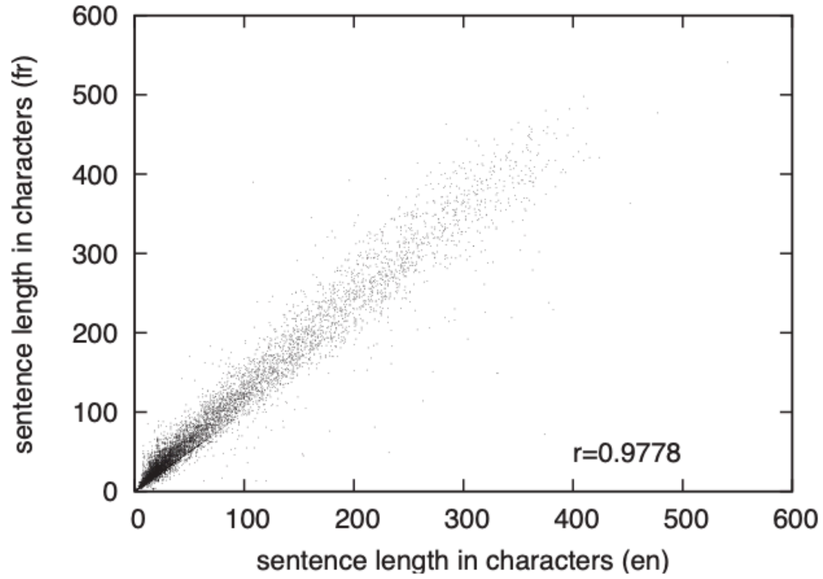
Lexical methods: assumption = corresponding sentences contain more corresponding words; use distribution of corresponding words in source and target language texts

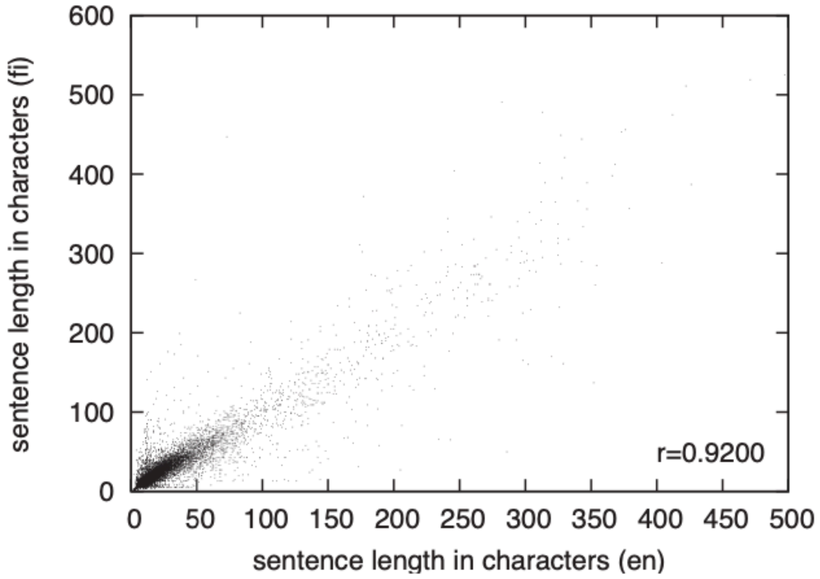
Combined methods: use lexical cues in length-based settings

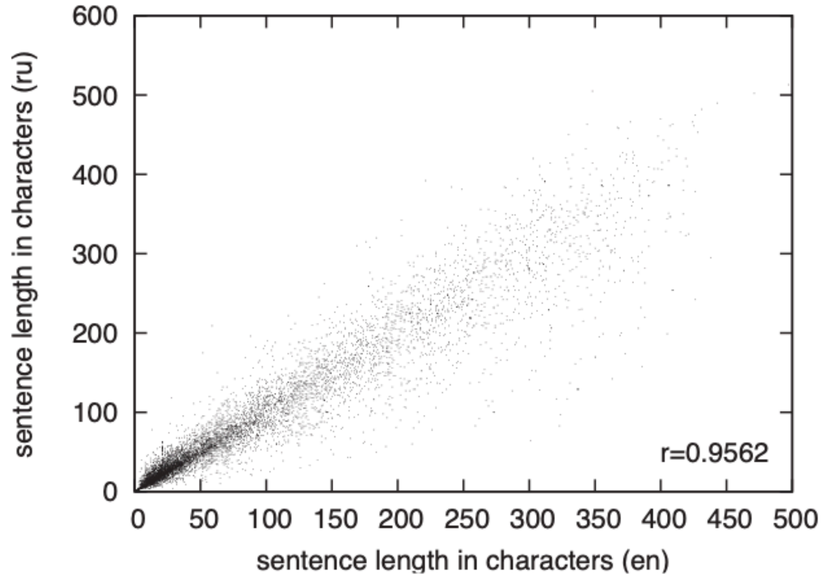
Length correlation (Gale & Church)

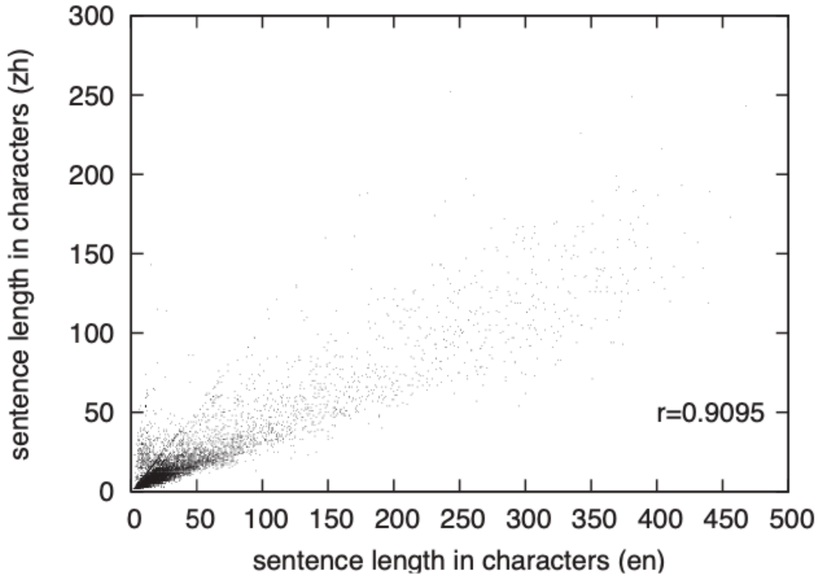


Length correlation other languages









String lengths

target language	ID
Vi var plågade av en epidemi som var mer hänsynslös än krig .	66
En pest som kom att döda mer än hälften av riket .	54
Vart kom den ifrån ?	22
Vem spred den ?	16
Prästerna berättade att det var Guds bestraffning .	54
För vilken synd ?	19
Nej , vi visste sanningen	26
Detta var inte Guds verk , utan djävulens .	45
Eller trolldom .	17
Vi var tvungna att besegra en demon .	38
För Guds förlåtelse .	25
Broder Thomas .	16

ID	source language
92	Eine Seuche , grausamer und erbarmungsloser als der Krieg , war über uns hereingebrochen .
70	Eine Pestseuche , die die Hälfte unseres Königreiches dahinraffte .
17	Wo kam sie her ?
27	Wie verbreitete sie sich ?
54	Die Priester sagten uns , sie sei die Strafe Gottes .
26	Aber für welche Sünde ?
73	Welches Gebot hatten wir gebrochen , dass wir so etwas verdient hatten ?
34	Nein , wir kannten die Wahrheit .
64	Das war nicht Gottes Werk , sondern Teufelei ... oder Hexerei .
86	Und unsere Aufgabe bestand darin , für die Heilung durch Gott den Dämon zu fangen .
16	Bruder Thomas !

Finding the optimal alignment

Define a cost function *cost(a)*

- penalize mismatch in string lengths
- penalize uncommon alignment types
- …

Search the alignments that minimizes the overall cost

$$A' = \underset{A}{argmin} \sum_{i} cost(a_i)$$

Use dynamic programming for efficient computation

The use of cognates

9	- Mr. Gibbs that will do !	- Mr Gibbs , det räcker .	6
10	- She was singing about pirates .	Hon sjöng om sjörövare .	7
11	Bad luck to be singing about pirates with us mired in this unnatural fog .	Att sjunga om sjörövare i denna dimma ger otur , om jag får säga det .	8
12	Mark my words .	Nu har ni fått säga det ...	9
13	Consider them marked .	- Iväg med er .	10
14	On your way .	- Ja , löjtnant .	11
15	Aye , Lieutenant .	Att ha kvinnor ombord ger också otur , även om hon är en miniatyr ...	12
16	It' s bad luck to have a woman on board , too , even a miniature one .	Att träffa en sjörövare vore spännande .	13
17	I think it' d be rather exciting to meet a pirate .		
18	Think again , Miss Swann .	Tänk efter , miss Swann .	14
19	Vile and dissolute creatures , the lot of them .	De är avskyvärda och lastbara allihop .	15
20	I intend to see to that any man who sails under a pirate flag or ... ... wears a pirate brand gets what he deserves .	Jag ska se till att alla som seglar under sjörövarflagg får vad de förtjänar :	16
21	A short drop and a sudden stop .	Kort fall med snabbt stopp ...	17
22	Lieutenant Norrington , I appreciate your fervor .	Löjtnant Norrington , jag förstår er iver -	18
23	But I' m ... I' m concerned about the effect this subject will have upon my daughter .	- men jag är orolig för hur ämnet påverkar min dotter .	19
24	My apologies , Governor Swann .	Jag ber om ursäkt , guvernör Swann .	20

Common Tools

hunalign

- combines length-based + lexical matching
- two-step option: induced lexicon

gargantua

- multipass procedure (length-based + induced lexicon)
- final two-step clustering approach

bleualign

- translate (MT) and align based on BLEU match

Common tools

LF aligner

- based on hunalign + interface
- can produce multiparallel data sets

srtalign

- time-based subtitle alignment
- synchronisation based on lexical matches

yalign

- trainable sentence aligner for comparable corpora

Take-home messages

Parallel corpora and neural MT

- essential supervision to train model parameters
- need to be sentence-aligned
- growing resources available (but still quite scarce)

Tools and approaches

- crawling, converting, lang-identification, pre-processing
- unsupervised sentence alignment (length, cognates, …)
- some level of noise is OK