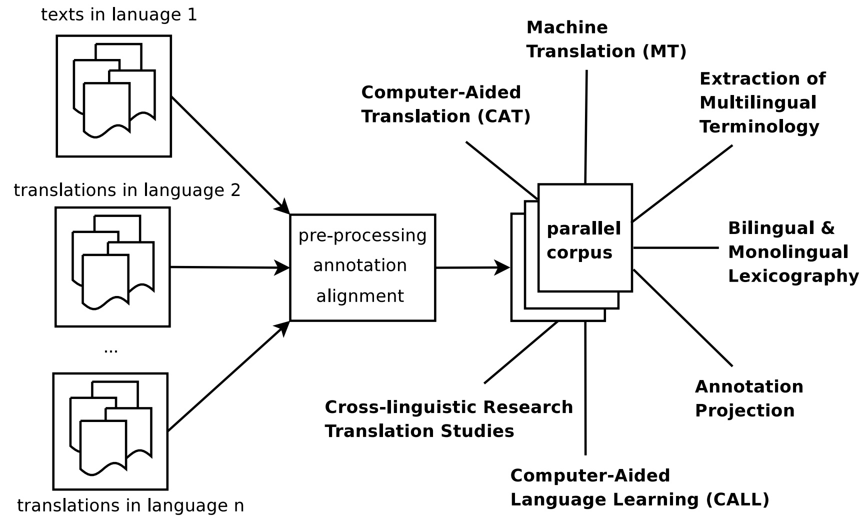
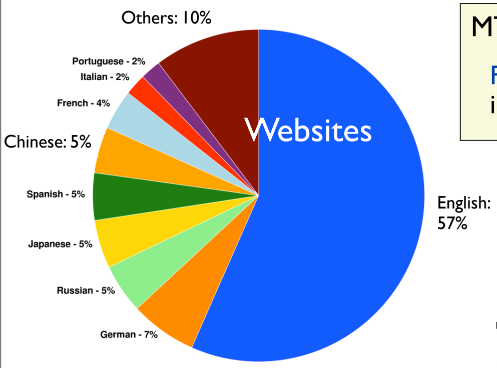
Building Parallel Corpora

Jörg Tiedemann
Uppsala University
jorg.tiedemann@lingfil.uu.se

Parallel Corpora Are Very Useful

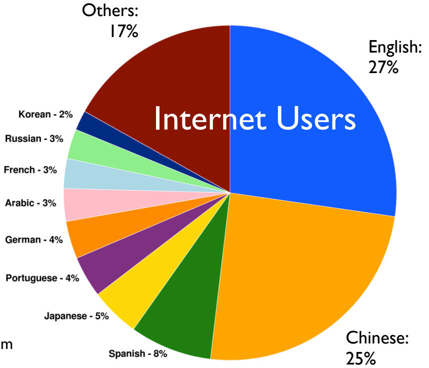


The Web is a Multi-Lingual Place!



MT for Gisting Purposes:

Find and understand
information in other languages

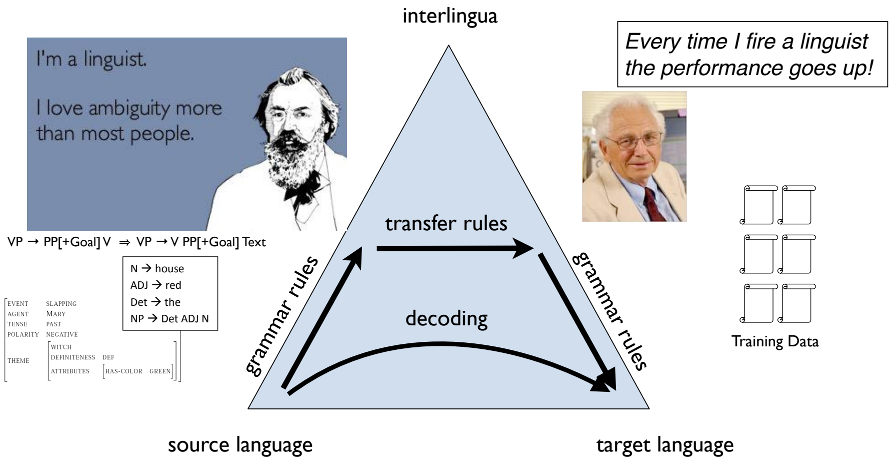


> 2 billion Internet users
> 12 billion indexed web pages

Sources: W3Techs.com, Internet World Stats, WorldWideWebSize.com

Rule-Based vs. Data-Driven MT

*Every time I fire a linguist
the performance goes up!*



Lack of MT Training Data is a Big Issue

Macedonian - English:

It's a simple matter of self-preservation.
It's simply a question of себесочувување.

Your girlfriend's very cynical.
Пријателката цинична you very much.

Catalan - English:

She just went through a breakup.
Just come in for a ruptura

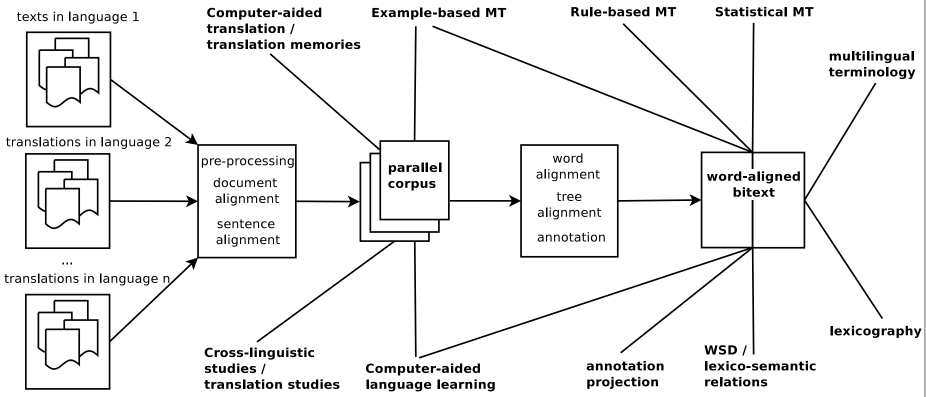
Galician - English:

I promise I will return it to you.
Cha devolverei I promise you that.

What is a Parallel Corpus?

English	French	German	Old French
original text 1	translation of text 1	translation of text 1	
translation of text 2	original text 2		
	translation of text 3	translation of text 3	translation of text 3

Building and Using Parallel Corpora



Building Parallel Corpora

find and collect translated documents

pre-processing

- conversion
- sentence boundary detection
- tokenization

alignment

- document alignment
- paragraph alignment (optional)
- sentence alignment

UPPSALA
UNIVERSITET

Step 1: Pre-processing

Example: Subtitles

```
1
00:00:26,500 --> 00:00:28,434
Spend all day with us.

2
00:00:28,502 --> 00:00:30,436
There are two-- pardon me--

3
00:00:30,504 --> 00:00:34,440
two of everything in
every Noah’s arcade.

4
00:00:34,508 --> 00:00:36,361
That means two of Zantar,

5
00:00:36,361 --> 00:00:36,884
That means two of Zantar,
```

```
<?xml version="1.0" encoding="utf-8"?>
<document>
  <s id="1">
    <time id="T1S" value="00:00:26,500" />
    <w id="1.1">Spend</w>
    <w id="1.2">all</w>
    <w id="1.3">day</w>
    <w id="1.4">with</w>
    <w id="1.5">us</w>
    <w id="1.6">.</w>
    <time id="T1E" value="00:00:28,434" />
  </s>
  <s id="2">
    <time id="T2S" value="00:00:28,502" />
    <w id="2.1">There</w>
    <w id="2.2">are</w>
    <w id="2.3">two</w>
    <w id="2.4">--</w>
    <w id="2.5">pardon</w>
    <w id="2.6">me</w>
    <w id="2.7">--</w>
    <time id="T2E" value="00:00:30,436" />
    <time id="T3S" value="00:00:30,504" />
    <w id="2.8">two</w>
    <w id="2.9">of</w>
    <w id="2.10">everything</w>
    ...
```

UPPSALA
UNIVERSITET

Step 2: Paragraph / Sentence Alignment

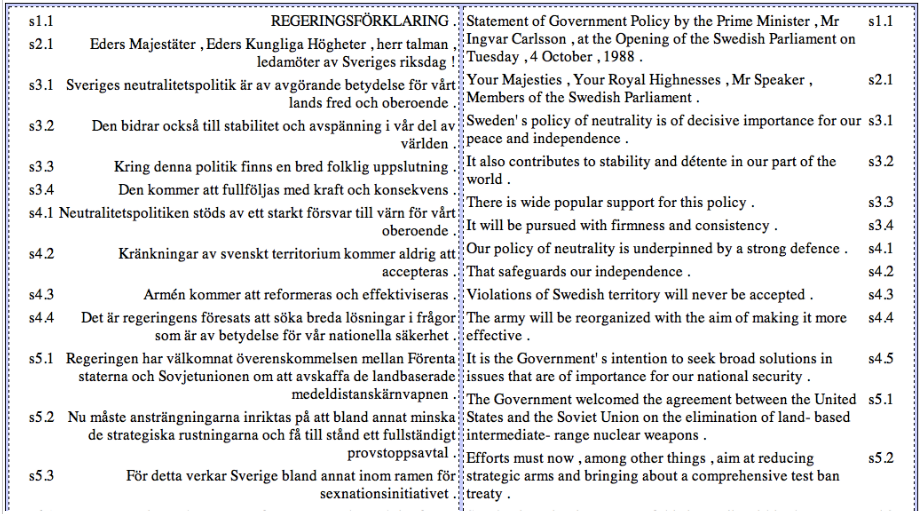
Task: align corresponding sentences to each other (may be sequences of sentences)

Assumption: sentence alignment can be done monotonically (no crossing links!)

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

UPPSALA
UNIVERSITET

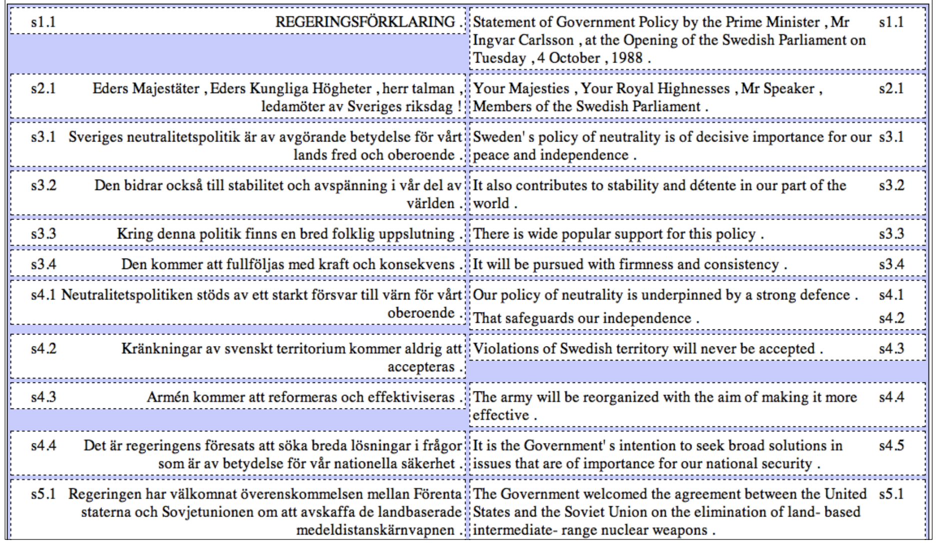
Simple Case: Legal Documents



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
s4.2	Kränkningar av svenskt territorium kommer aldrig att accepteras .	That safeguards our independence .	s4.2
s4.3	Armén kommer att reformeras och effektiviseras .	Violations of Swedish territory will never be accepted .	s4.3
s4.4	Det är regeringens föresats att söka breda lösningar i frågor som är av betydelse för vår nationella säkerhet .	The army will be reorganized with the aim of making it more effective .	s4.4
s5.1	Regeringen har välkomnat överenskommelsen mellan Förenta staterna och Sovjetunionen om att avskaffa de landbaserade medeldistanskärnvapnen .	It is the Government' s intention to seek broad solutions in issues that are of importance for our national security .	s4.5
s5.2	Nu måste ansträngningarna inriktas på att bland annat minska de strategiska rustningarna och få till stånd ett fullständigt provstoppsavtal .	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
s5.3	För detta verkar Sverige bland annat inom ramen för sexnationsinitiativet .	Efforts must now , among other things , aim at reducing strategic arms and bringing about a comprehensive test ban treaty .	s5.2

UPPSALA
UNIVERSITET

Simple Case: Legal Documents



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 . That safeguards our independence .	s4.1 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

Hard Case: Movie Subtitles

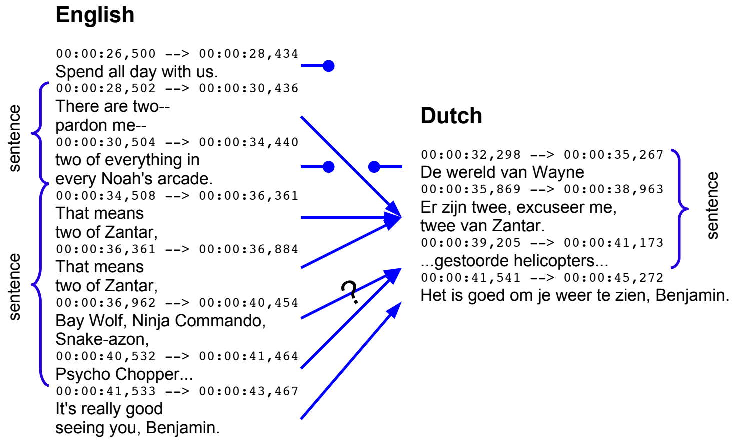
English

00:00:26,500 --> 00:00:28,434
Spend all day with us.
00:00:28,502 --> 00:00:30,436
There are two--
pardon me--
00:00:30,504 --> 00:00:34,440
two of everything in
every Noah's arcade.
00:00:34,508 --> 00:00:36,361
That means
two of Zantar,
00:00:36,361 --> 00:00:36,884
That means
two of Zantar,
00:00:36,962 --> 00:00:40,454
Bay Wolf, Ninja Commando,
Snake-azon,
00:00:40,532 --> 00:00:41,464
Psycho Chopper...
00:00:41,533 --> 00:00:43,467
It's really good
seeing you, Benjamin.

Dutch

00:00:32,298 --> 00:00:35,267
De wereld van Wayne
00:00:35,869 --> 00:00:38,963
Er zijn twee, excuseer me,
twee van Zantar.
00:00:39,205 --> 00:00:41,173
...gestoorde helicopters...
00:00:41,541 --> 00:00:45,272
Het is goed om je weer te zien,
Benjamin.

Hard Case: Movie Subtitles



English

00:00:26,500 --> 00:00:28,434
Spend all day with us.
00:00:28,502 --> 00:00:30,436
There are two--
pardon me--
00:00:30,504 --> 00:00:34,440
two of everything in
every Noah's arcade.
00:00:34,508 --> 00:00:36,361
That means
two of Zantar,
00:00:36,361 --> 00:00:36,884
That means
two of Zantar,
00:00:36,962 --> 00:00:40,454
Bay Wolf, Ninja Commando,
Snake-azon,
00:00:40,532 --> 00:00:41,464
Psycho Chopper...
00:00:41,533 --> 00:00:43,467
It's really good
seeing you, Benjamin.

Dutch

00:00:32,298 --> 00:00:35,267
De wereld van Wayne
00:00:35,869 --> 00:00:38,963
Er zijn twee, excuseer me,
twee van Zantar.
00:00:39,205 --> 00:00:41,173
...gestoorde helicopters...
00:00:41,541 --> 00:00:45,272
Het is goed om je weer te zien, Benjamin.

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 Huuit .	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 teooohio	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

UPPSALA UNIVERSITET

It was actually German and Swedish

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

UPPSALA UNIVERSITET

Sentence Alignment Approaches

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)

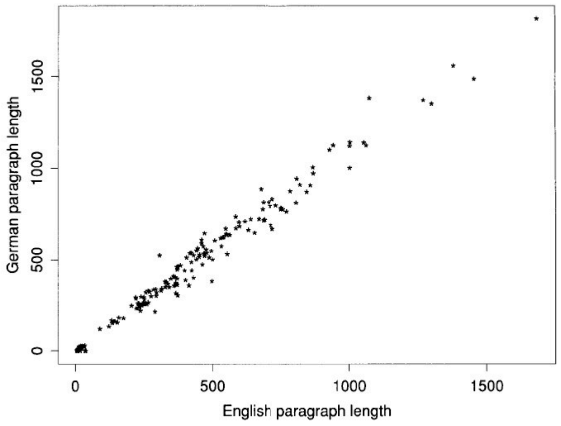
Lexical methods: assumption = corresponding sentences contain more corresponding words; use distribution of corresponding words in source and target language texts

Combined methods: use lexical cues in length-based settings

UPPSALA UNIVERSITET

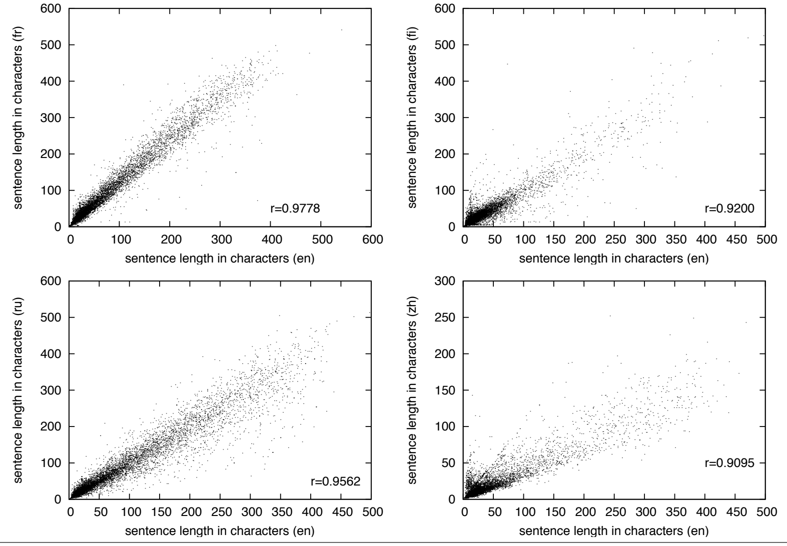
Gale & Church: A Length-Based Model

Strong correlation between paragraph lengths



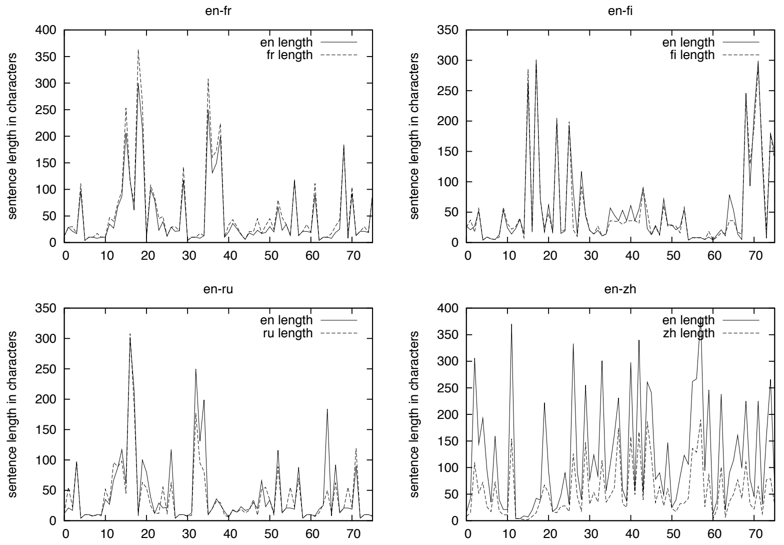
UPPSALA UNIVERSITET

Translated KDE System Messages



UPPSALA UNIVERSITET

Sentence Lengths as Alignment Signals



UPPSALA UNIVERSITET

Compare 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 !

UPPSALA UNIVERSITET

The Gale & Church Approach

Generative Model

- each source character generates a target character

Minimize alignment costs

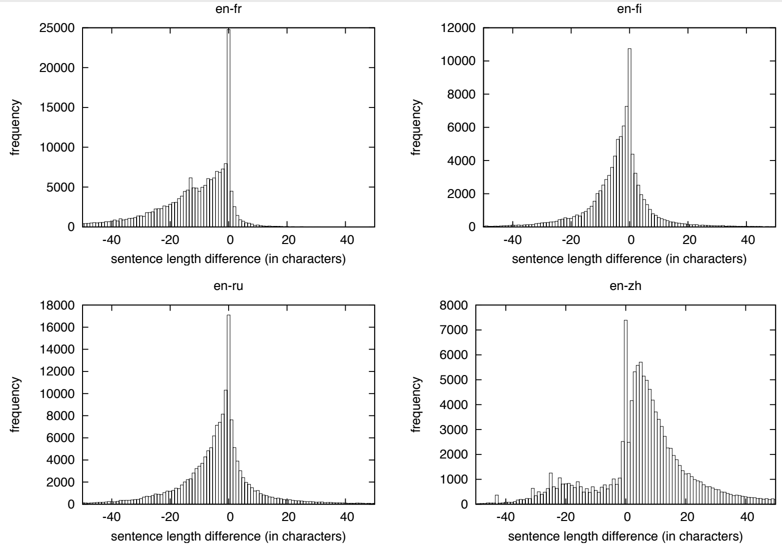
- minimize length difference
- uncommon alignment types have high costs

Dynamic programming solution

- fixed set of alignment types (0:1, 1:1, 1:2, 2:2)
- fixed cost parameters
- finds global optimum (for given setup)

UPPSALA UNIVERSITET

Sentence Length Difference (KDE)



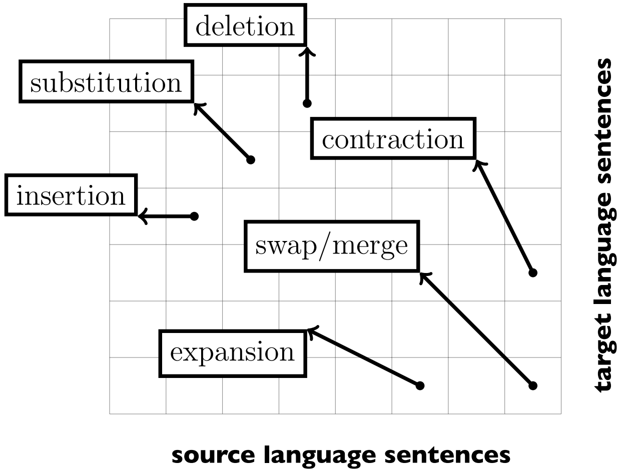
Alignment Type Probabilities (Prior)

Estimated from an aligned parallel corpus:

- P(type = 1:1) = 0.89 (substitution)
- P(type = 1:0) = 0.0099/2 (insertion)
- P(type = 0:1) = 0.0099/2 (deletion)
- P(type = 2:1) = 0.0891/2 (expansion)
- P(type = 1:2) = 0.0891/2 (contraction)
- P(type = 2:2) = 0.011 (merging/swap)

Simple idea: Keep them fixed!

The Gale & Church Approach



deletion

substitution

contraction

insertion

swap/merge

expansion

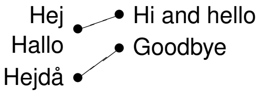
target language sentences

source language sentences

The Gale & Church Approach: Example

			Hej		Hallo		Hejdå	
	0		$6.37^{1:0}$ – –	– – –	$13.16^{1:0}$ – –	– – –	$20.17^{1:0}$ – –	– – –
Hi and hello	– – $8.12^{0:1}$	– –	$14.48^{1:0}$ $1.69^{1:1}$ $14.48^{0:1}$	– – –	$8.49^{1:0}$ $7.51^{1:1}$ $21.28^{0:1}$	**$3.58^{2:1}$** – –	$10.58^{1:0}$ $14.09^{1:1}$ $28.28^{0:1}$	$9.57^{2:1}$ – –
Goodbye	– – $15.31^{0:1}$	– – –	$21.68^{1:0}$ $8.94^{1:1}$ $8.89^{0:1}$	– – $5.86^{1:2}$	$12.65^{1:0}$ $2.09^{1:1}$ $10.77^{0:1}$	$11.35^{2:1}$ $5.89^{2:2}$ $11.59^{1:2}$	$9.09^{1:0}$ **$3.82^{1:1}$** $16.77^{0:1}$	$5.30^{2:1}$ $11.72^{2:2}$ $18.12^{1:2}$

Resulting alignment (one 2-to-1 match & one 1-to-1 match):



Hej

Hallo

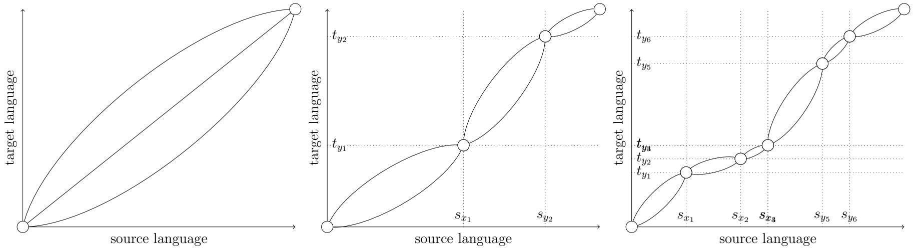
Hejdå

Hi and hello

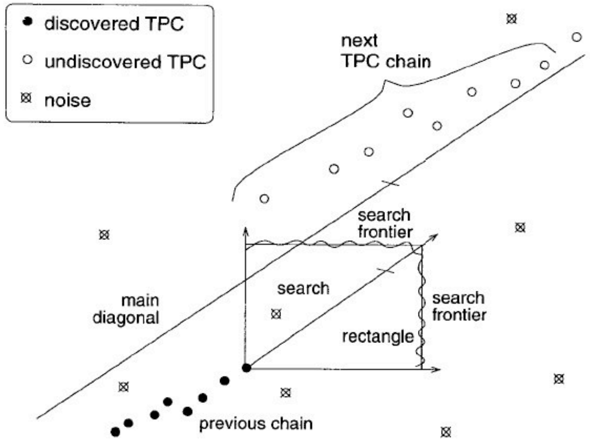
Goodbye

Idea 2: Lexical Matching Approaches

- discover translation equivalents (close to diagonal)
- align sentences that do not violate lexical links (a lot)



Lexical Translation Chains (Melamed)



Sentence Alignment Quality

How good is my alignment algorithm?

Depends very much on the input data

- complete (mostly literal) translations
- no inserted text
- correct paragraph / sentence boundaries

Error propagation (especially length-based approaches)

Other clues:

- formatting
- time information (movie subtitles)

Summary on Building Parallel Corpora

Collect and convert translated documents

Parallel corpora need alignment

- paragraph alignment
- sentence alignment

Automatic sentence alignment

- length-based approaches
- lexical matching approaches

Problems: incomplete translations, noisy texts, ...

Beyond Sentence Alignment

Is it possible to find links between smaller units than paragraphs and sentences?

UPPSALA
UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

(Knight, 1997)

UPPSALA
UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

UPPSALA
UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

UPPSALA
UNIVERSITET

Translate Centauri into Arcturan

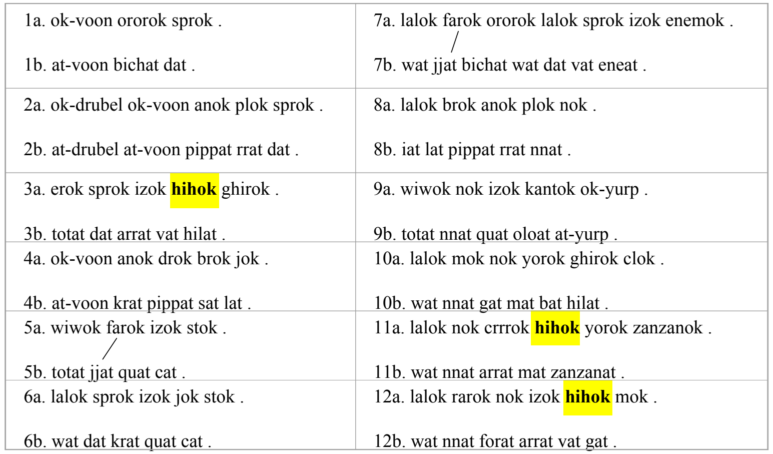
Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok **crrrok** hihok yorok zanzanok . ??? 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

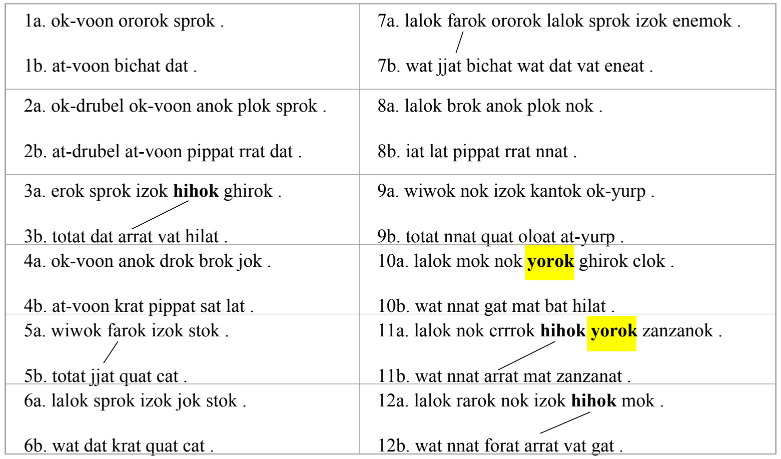


1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

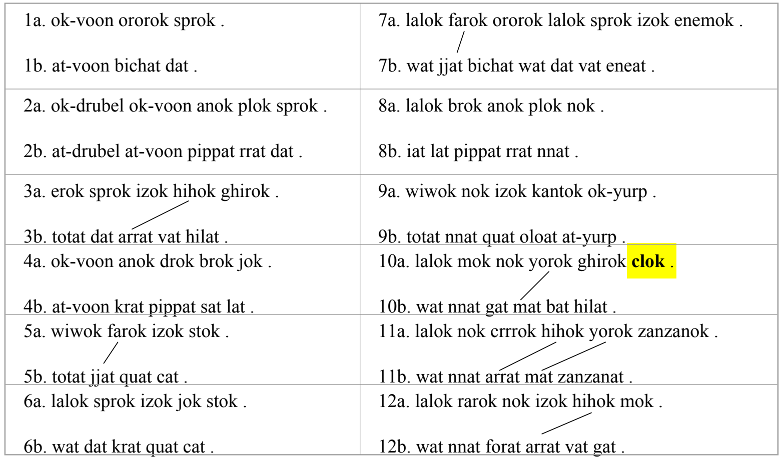


1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok **yorok** ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** **yorok** zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

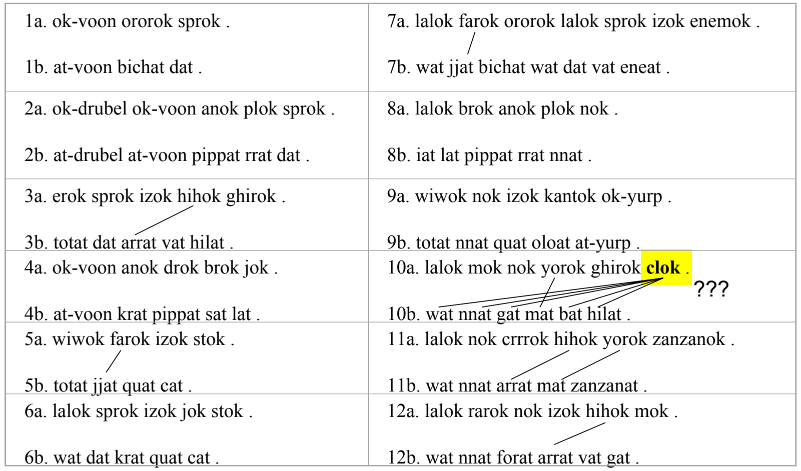


1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

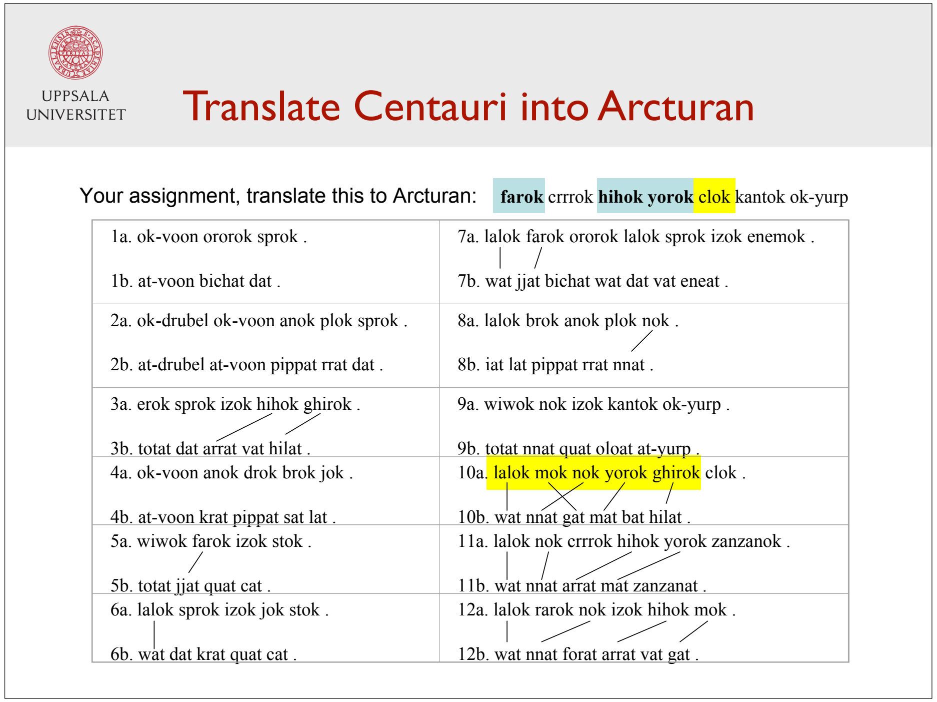
UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp



1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . ??? 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

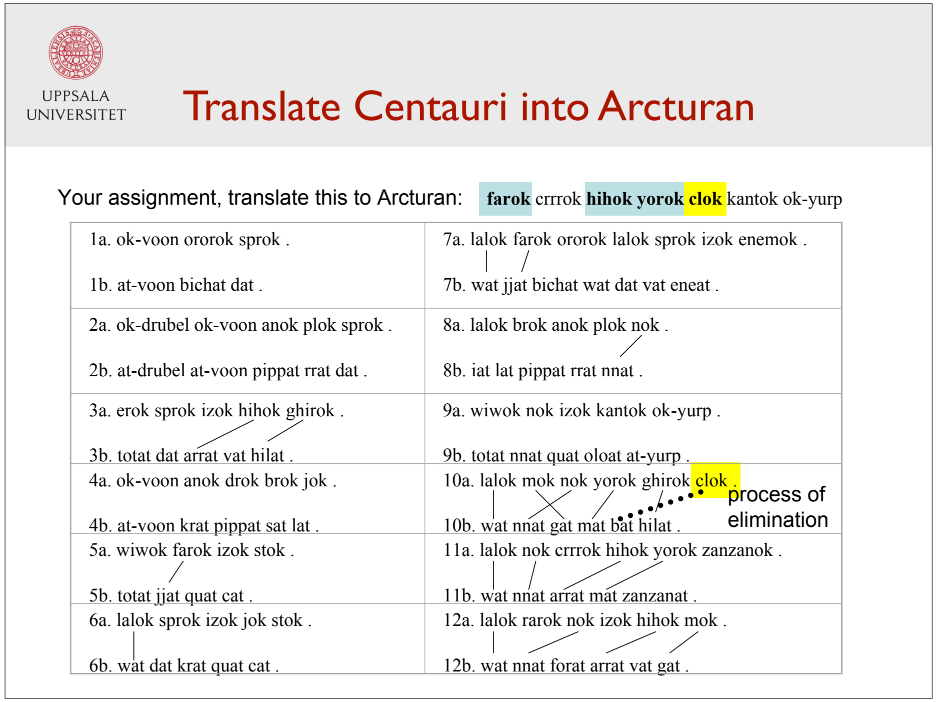


UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .



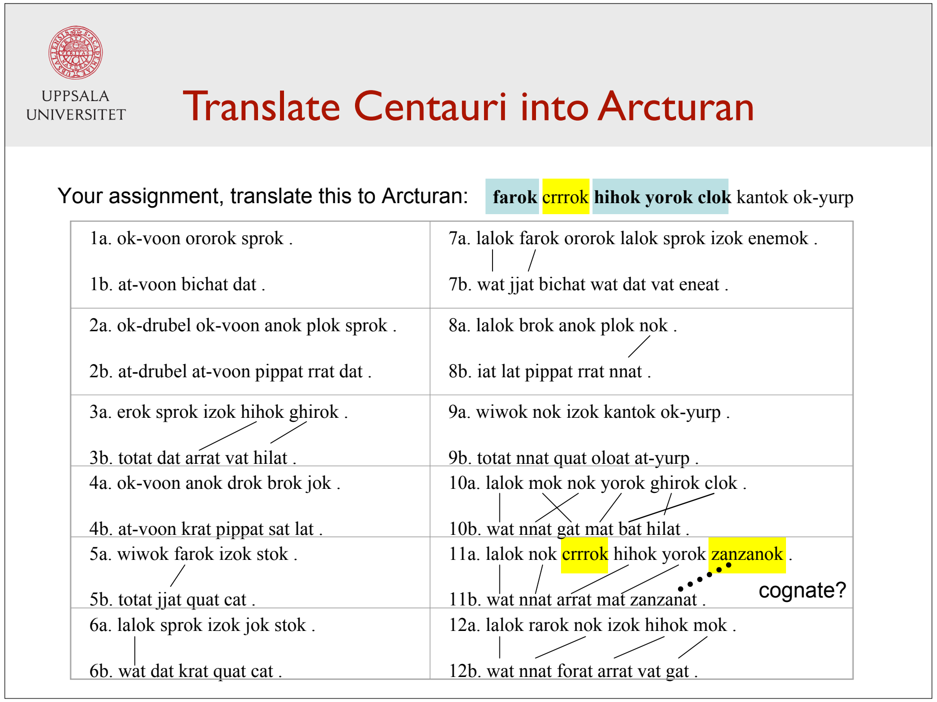
UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

process of elimination



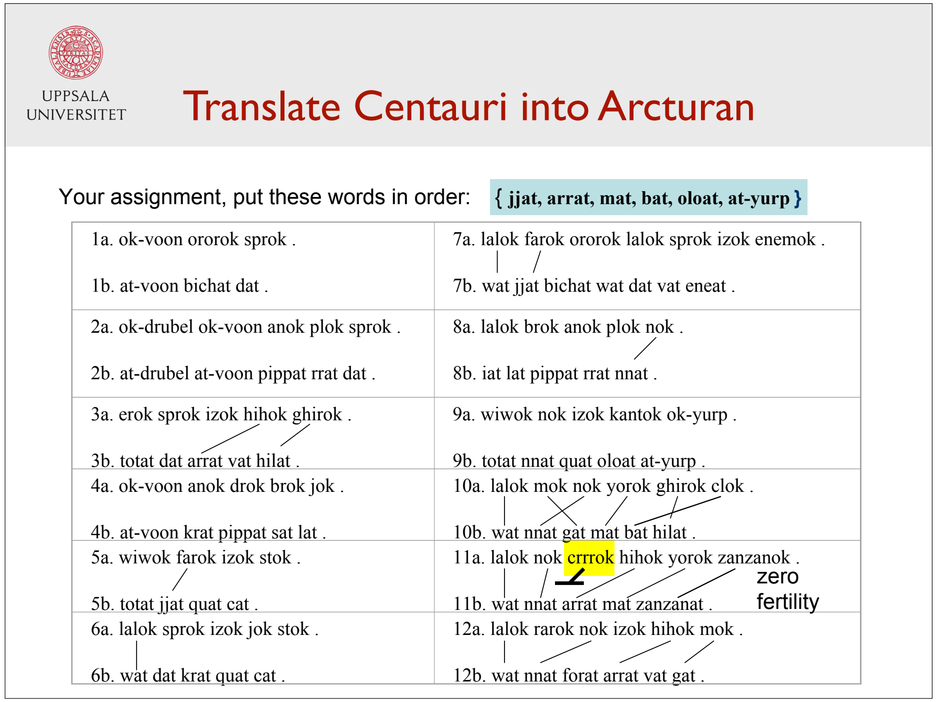
UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

cognate?



UPPSALA UNIVERSITET

Translate Centauri into Arcturan

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

zero fertility

Translate Centauri into Arcturan

Clients do not sell pharmaceuticals in Europe => **Clientes no venden medicinas en Europa**

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Conclusions

It is possible to find links between words

... without any prior knowledge

- Distribution of words is a good clue
- Orthographic similarities may be another
- Sentence alignment is essential to make this work!

Word alignment is useful for many things

- learn to translate
- lexicography
- annotation projection
- ...

Questions?