CORPUS

Parallel Corpora for Everyone
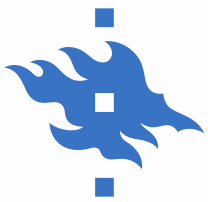
Jörg Tiedemann

Department of Modern Languages

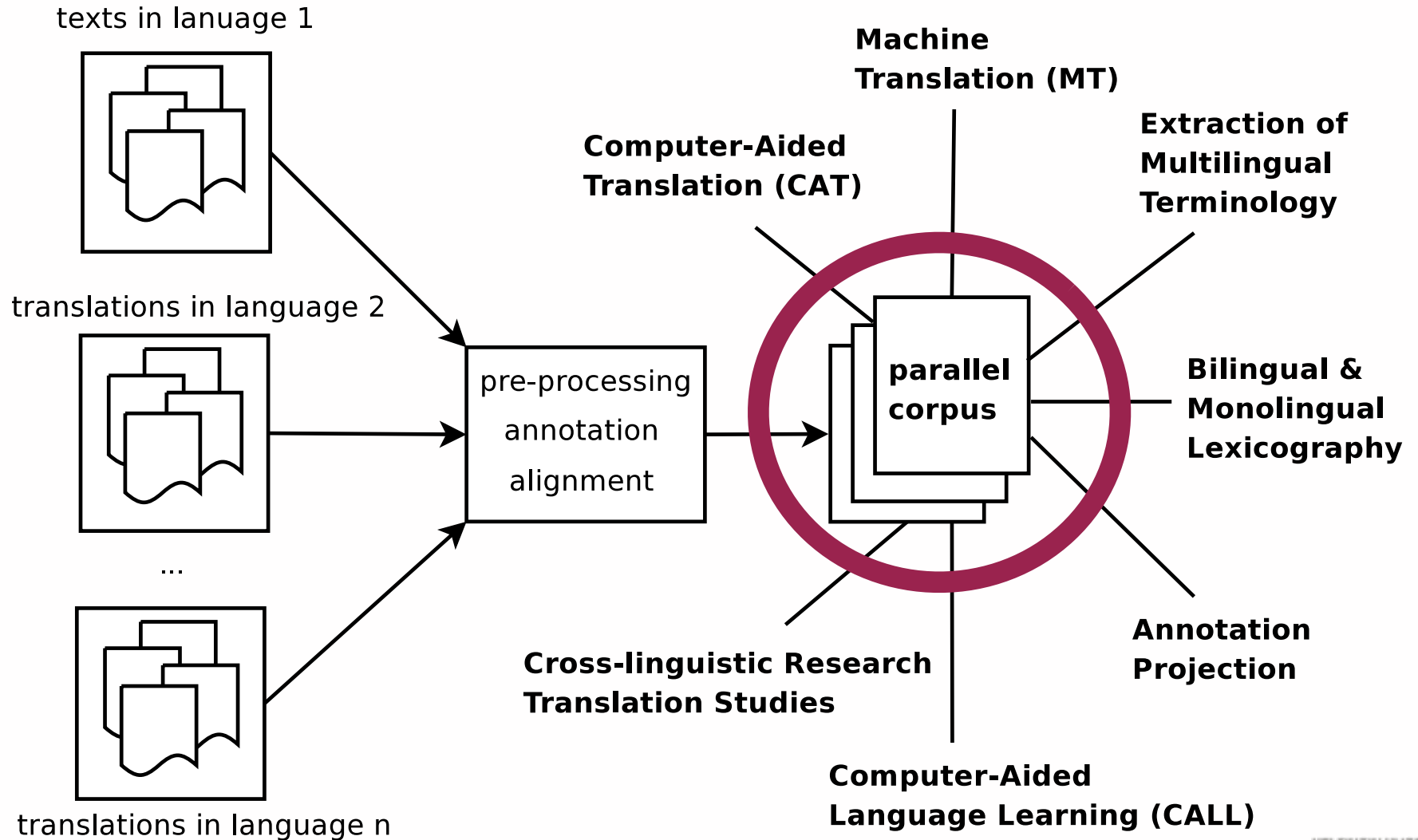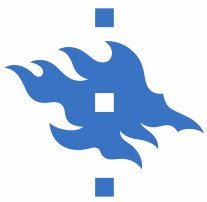**University of Helsinki**

Rosendal 2002

The Amazing Utility of Parallel Corpora

texts in lanuage 1

translations in language 2

...

translations in language n

pre-processing

annotation

alignment

parallel corpus

Machine Translation (MT)

Computer-Aided Translation (CAT)

Extraction of Multilingual Terminology

Bilingual & Monolingual Lexicography

Annotation Projection

Cross-linguistic Research Translation Studies

Computer-Aided Language Learning (CALL)

OPUS - an open source parallel corpus
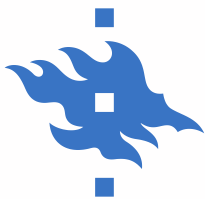
http://logos.uio.no/opus/

Jörg Tiedemann
Department of Linguistics
Uppsala University
Box 527
SE-751 20 Uppsala, Sweden
joerg@stp.ling.uu.se

Lars Nygaard
Tekstlaboratoriet HF
University of Oslo
Postboks 1102 Blindern
0317 Oslo
lars.nygaard@ilf.uio.no

1   Introduction

Parallel corpora are useful in a wide variety of research areas, particularly in machine translation and lexicography. However, parallel corpora have been few, often unrepresentative, and not generally available. The aim of the OPUS project is to provide a public collection of parallel corpora which can be freely used and distributed. This makes it possible for everyone to

Now at http://opus.lingfil.uu.se

OᵣPUS ... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@lingfil.uu.se >

Latest News

- 2016-01-08: New version: OpenSubtitles2016
- 2015-10-15: New versions of TED2013, NCv9
- 2014-10-24: New: JRC-Acquis
- 2014-10-20: NCv9, TED talks, DGT, WMT
- 2014-08-21: New: Ubuntu, GNOME
- 2014-07-30: New: Translated Books
- 2014-07-27: New: DOGC, Tanzil
- 2014-05-07: Parallel coref corpus ParCor

Search & download resources: -- select -- -- select -- all

Search & Browse

- OPUS multilingual search interface
- Europarl v7 search interface
- Europarl v3 search interface
- OpenSubtitles search interface
- EUconst search interface
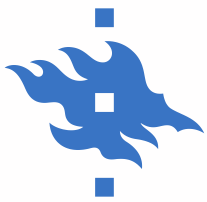- Word Alignment Database

Tools & Info

- OPUS Wiki
- Tools for tagging and parsing
- Downloads (tools and models)
- Other annotation and corpus tools
- Experimental visualization tool for monolingual and parallel treebanks (demo)
- Uplug at bitbucket
- A reliable Language Identifier

Sub-corpora (downloads & infos):

- Books - A collection of translated literature (DOGC2014-07-17.tar.gz - 236 MB)
- DGT - A collection of EU Translation Memories provided by the JRC
- DOGC - Documents from the Catalan Goverment (DOGC2014-07-17.tar.gz - 702 MB)
- ECB - European Central Bank corpus
- EMEA - European Medicines Agency documents (EMEA0.3.tar.gz - 5.0 GB)
- The EU bookshop corpus (EUbookshop/EUbookshop0.2.tar.gz - 33 GB)
- EUconst - The European constitution (EUconst0.1.tar.gz - 67 MB)
- EUROPARL v7 - European Parliament Proceedings (Europarlv7.tar.gz - 8.4 GB)
- EUROPARL - European Parliament Proceedings (Europarl3.tar.gz - 3.6 GB)
- GNOME - GNOME localization files (GNOME2014-08-20.tar.gz - 9 GB)
- Global Voices - News stories in various languages (GlobalVoices2015.tar.gz - 1.1 GB)
- The Croatian - English WaC corpus (hrenWaC1.tar.gz - 48 MB)
- JRC-Acquis- legislative EU texts
- KDE4 - KDE4 localization files (v.2) (KDE4.tar.gz - 1.4 GB)
- KDEdoc - the KDE manual corpus (KDEdoc.tar.gz - 35 MB)
- MBS - Belgisch Staatsblad corpus
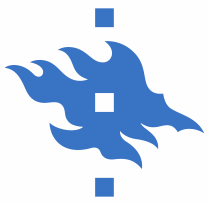- MultiUN - Translated UN documents

OrPUS

Highlights

- over 200 language and language variants
- 29 million documents, 3.2 billion sentences, 28 billion tokens
- several domains (legislation, medical, subtitles ...)
- 12,572 language pairs, 10.8 billion translation units
- available in several formats (OPUS XML, TMX, Moses)

Tools & online services

- tools for conversion, annotation, alignment
- online search interfaces
- word alignment database + concordances

http://opus.lingfil.uu.se

What Is Included?

- Copyright-free Books
- DGT translation memories
- DOGC (Catalan Government)
- European Central Bank
- European Medicines Agency
- EU Bookshop
- EuroParl
- GNOME, KDE, OpenOffice, PHP, Ubuntu localisation files
- Global Voices News
- Croatian-English WaC

- JRC-Acquis
- Belgisch Stadsblad
- United Nations corpora
- News Commentary, WMT sets
- OpenSubtitles
- SETIMES
- Tatoeba
- TedTalks
- Tanzil Quran Translations
- Wikipedia, WikiSource

What Is It Used For?

**ACL 2016
FIRST CONFERENCE ON
MACHINE TRANSLATION (WMT16)**
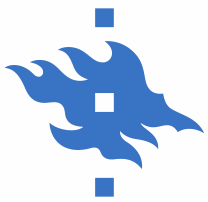
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

Reverso Context
The best way to translate words and expressions

English↔German ▸

Search for conversations in movies

Computer-Aided Language Learning

ЯR eLearnRussian.com

Home > Movie Central > Трое из Простоквашино

Трое из Простоквашино :

Part 1 :

English	Russian
CAT: You're eating your sandwich the wrong way, Uncle Fedor	Неправильно ты, дядя ешь.
You're holding it with the meat facing up, but you're supposed to have it the meat facing down.	Ты его колбасой к ве надо колбасой на язь
It will come out tastier that way!	Так вкуснее получится!
And how do you know that my name is Uncle Fedor?	А откуда ты знаешь, что меня дядей Фёдором зовут?
I know everyone in our building.	Я в нашем доме всех знаю,
I live up in the attic, and I can see everything	я на чердаке живу и мне всё видно.
Matroskin the Cat is my name. It's something of a family name	Кот Матроскин меня зовут, эта фамилия такая.
Currently they're remodeling the attic, so I have nowhere to live	Только сейчас наш чердак ремонтируют - мне жить негде.

http://elearnrussian.com

Download Resources

Search & download resources: fi (Finnish) ▼   lv (Latvian) ▼   all ▼

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, true = truecase

corpus	doc's	sent's	src tokens	trg tokens	XCES/XML	raw	TMX	Moses	mono	raw
DGT	26534	3.0M	46.3M	51.8M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
Europarl	7993	0.6M	12.2M	13.8M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
EMEA	1916	1.1M	12.2M	13.4M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
EUbookshop	1108	0.4M	12.0M	13.6M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
GNOME	1239	0.6M	2.3M	2.8M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
OpenSubtitles2016	286	0.4M	1.9M	2.2M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
ECB	1	56.1k	1.0M	1.0M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
JRC-Acquis	198	20.1k	0.7M	1.1M	[xces fi lv]	[fi lv]	[tmx]	[moses]		
OpenSubtitles2013	106	0.1M	0.7M	0.8M	[xces fi lv]	[fi lv]		[moses]		fi lv
OpenSubtitles2012	102	0.1M	0.7M	0.8M	[xces fi lv]	[fi lv]		[moses]		fi lv
KDE4	639	82.9k	0.4M	0.5M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
Ubuntu	290	68.3k	0.3M	0.2M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
EUconst	47	10.3k	0.1M	0.1M	[xces fi lv]	[fi lv]	[tmx]	[moses]	fi lv	fi lv
total	**40459**	**6.6M**	**90.8M**	**102.2M**	**6.6M**		**3.9M**	**5.8M**		

native XML and
standoff sentence alignment

MT training data

Download Resources

vian)   all

e data! (raw = untokenized, true = truecaser model, TM = phrase-based translation model)

on-line search

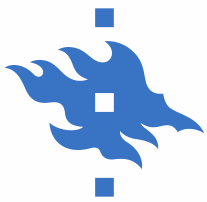| raw	TMX	Moses	mono	raw	true	TM	dic	freq			Browse Files
[fi lv]	[tmx]	[moses]	fi lv	fi lv	fi lv	fi-lv		fi lv		[sample]	[xml/fi][xml/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv	fi lv	fi-lv	dic	fi lv	[query]	[sample]	[xml/fi][xml/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv	fi lv	fi-lv	dic	fi lv	[query]	[sample]	[xml/fi][xml/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv	fi lv	fi-lv	dic	fi lv		[sample]	[xml/fi][xml/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv						[sample]	[xml/fi][xml/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv					[query]	[sample]	[xml/fi][xml/lv][raw/fi][raw/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv	fi lv	fi-lv	dic	fi lv	[query]	[sample]	[xml/fi][xml/lv][xml/fi-lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv						[sample]	[xml/fi][xml/lv]
[fi lv]		[moses]		fi lv			dic	fi lv		[sample]	
[fi lv]		[moses]		fi lv	fi lv	fi-lv	dic	fi lv		[sample]	
[fi lv]	[tmx]	[moses]	fi lv	fi lv	fi lv	fi-lv	dic	fi lv	[query]	[sample]	[xml/fi][xml/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv						[sample]	[xml/fi][xml/lv]
[fi lv]	[tmx]	[moses]	fi lv	fi lv	fi lv	fi-lv	dic	fi lv	[query]	[sample]	[xml/fi][xml/lv][xml/fi-lv]
	3.9M	5.8M									

monolingual data (raw text)

statistical translation models

word alignment dictionaries

individual files

On-Line Search

On-Line Search

"as" [tnt="JJ.*"] "as" <chunk_type="NP"> []+ </chunk>

sort = | unsorted |

| Reset Form |

| Run Query | | Distribution | | Frequencies |

Display: tokens = | word | phrases = | off | context = | sentence |

Keyword: ☐ | nearest | [_____] within | | | 5 words | from | match | ☐ inclus

Alignments: ☐ da ☑ de ☐ el ☐ es ☐ fi ☐ fr ☐ it ☐ nl ☐ pt ☑ sv

⏮ ⏪ | 1 – 20 | (Go) ⏩ ⏭   **[2305 matches]**   tokens =

1. Chapter 4, Koch (DE)	**de**
context We would like to ensure that there is a reference to this **as early as the recitals** and that the period within which the Council has to make a decision - which is not clearly worded - is set at a maximum of three months .	Wir möchten sicherstellen , daß hierauf bereits in den Erwägungsgründen hingewiesen wird und die uneindeutig formulierte Frist , innerhalb der der Rat eine Entscheidung treffen muß , auf maximal drei Monate fixiert wird .
2. Chapter 3, Färm (SV)	**de**
context Our experience of modern administration tells us that openness , decentralisation of responsibility and qualified evaluation are often **as effective as detailed bureaucratic supervision** .	Unsere Erfahrungen mit moderner Verwaltung besagen , daß Transparenz , Dezentralisation der Verantwortlichkeiten und eine qualifizierte Auswertung oft ebenso effektiv sind wie bürokratische Detailkontrolle .
3. Chapter 4, Caudron (FR)	**de**
context Essentially , my opinion is as follows : I deplore the fact that , in a field **as crucial as the transport** of dangerous goods , we keep putting off the adoption of harmonised standards to the detriment of human	Es läßt sich folgendermaßen zusammenfassen : ich bedaure , daß wir in einem so wichtigen Bereich wie der Beförderung gefährlicher Güter die Annahme von harmonisierten Normen auf Kosten

A Multilingual Word-Alignment Database

bul / chi / cze / dan / dut / ell / **eng** / est / fin / fre / ger / gle / heb / hrv / hun / ice / ita / jpn / lav / lit / mlt / nor / pob / pol / por / rum / rus / slo / slv / spa / swe / t

OPUS: Search Word Alignment Database for *eng*

☐ bul ☐ chi ☐ cze ☐ dan ☑ dut ☑ ell ☐ est ☐ fin
☑ fre ☐ ger ☐ gle ☐ heb ☐ hrv ☐ hun ☐ ice ☑ ita
☐ jpn ☐ lav ☐ lit ☐ mlt ☐ nor ☐ pob ☐ pol ☑ por
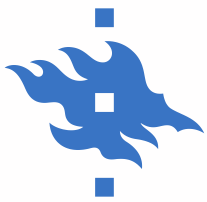☐ rum ☐ rus ☐ slo ☐ slv ☑ spa ☑ swe ☑ tur

- results from automatic word alignment
- wildcard symbols '%' and '_' allowed
- click on translations to query these words with their alignm
- click on frequencies to get concordance lines from the cor
- the concordancer does not use word alignment

honey (select) from ⦿ all ○ EUconst ○ Europarl3 ○ OpenSubtitles

dut	>>
663 honing	✓✗
180 schat	✓✗
108 schatje	✓✗
95 lieverd	✓✗
42 liefje	✓✗

ell	>>
379 μέλι	✓✗
279 μελιού	✓✗
45 το μέλι	✓✗
11 αγάπη	✓✗
9 γλυκιά	✓✗

fre	>>
697 miel	✓✗
76 chérie	✓✗
54 le miel	✓✗
42 miels	✓✗
37 chéri	✓✗

ita	>>
735 miele	✓✗
41 il miele	✓✗
21 del miele	✓✗
6 tesoro	✓✗
6 cara	✓✗

por	>>
759 mel	✓✗
271 querida	✓✗
130 querido	✓✗
28 o mel	✓✗
26 Querida	✓✗

spa	>>
813 miel	✓✗
369 cariño	✓✗
64 querida	✓✗
55 mieles	✓✗
42 querido	✓✗

swe	>>
541 honung	✓✗
245 älskling	✓✗
57 honungen	✓✗
30 raring	✓✗
26 gumman	✓✗

tur	>>
220 tatlım	✓✗
29 hayatım	✓✗
17 canım	✓✗
14 Tatlım	✓✗
5 Hayatım	✓✗

Examples from the **OpenSubtitles** corpus

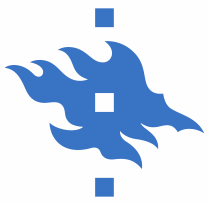| 901795	Show me the way home , **honey** .
swe	Visa mig vägen hem , **älskling** .
2228799	What's up , **honey** ?
swe	Vad är det , **älskling** ?
2230747	Clara , sweetie , **honey** .
swe	Clara , **älskling** ...
2230765	Clara , sweetie , **honey** .
swe	Clara , **älskling** ...
7905885	Let's go back , **honey** .
swe	Vi åker tillbaka , **älskling** !
7957671	Jerry , **honey** ?
swe	Jerry , **älskling** ...
9920396	Kenai , **honey** , shh , shh , shh , shh , shh .
swe	Kenai , **älskling** .
10083255	Here we go , **honey** .
swe	Nu gäller det **älskling** .
10730936	Well , I just wanted to get a little closer to you , **honey** .
swe	Jag ville bara komma lite närmare dig , **älskling** .

Latest News

Extended version of movie subtitle corpus

- 2.8 million subtitle files in 60 languages with a total of over 17 billion tokens in 2.6 billion sentences and sentence fragments

- All languages are sentence-aligned: 1,689 bitexts

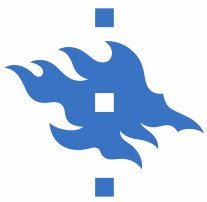- Release year, link to IMDb, genre, time info

Intra-lingual alignments

- Alternative translations of movie subtitles

- Classified into different categories

Example: English - Swedish

```
<s id="100">
    <time id="T82S" value="00:05:19,453" />
That sounds crazy.
    <time id="T82E" value="00:05:20,853" />
  </s>
  <s id="101">
    <time id="T83S" value="00:05:20,854" />
It felt like I could hear a song.
    <time id="T83E" value="00:05:23,944" />
  </s>
  <s id="102">
    <time id="T84S" value="00:05:23,991" />
(HUMMlNG 4-NOTE TUNE)
    <time id="T84E" value="00:05:26,491" />
  </s>
  <s id="103">
    <time id="T85S" value="00:05:31,665" />
See what you can find on that progression.
  </s>
  <s id="104">
It's, um... (breathing heavily)
    <time id="T86E" value="00:05:36,535" />
  </s>
```

```
  <s id="77">
    <time id="T55S" value="00:05:19,198" />
Det här låter tokigt.
  </s>
  <s id="78">
Jag hörde en melodi.
    <time id="T55E" value="00:05:22,827" />
  </s>
  <s id="79">
    <time id="T56S" value="00:05:31,711" />
Se vad du kan hitta om den tonföljden.
    <time id="T56E" value="00:05:36,257" />
  </s>
```
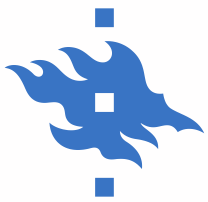
Alternative Subtitles

Spelling Errors / Variation:

```
Only Magneto is capabl of something like this .
Only Magneto is capable of something like this .

I accuse those who are asleep ...
l accuse those whoo are asleep ..

And the visionary told them many other things .
And the visionary told the any other things .

However her heart swayed , Edith suffered .
However her heart swayed , Edih suffered .
```

Alternative Subtitles

Insertions:

```
EMMA :
What exactly is this Sasaki Academy ?
What exactly is this Sasaki Academy ?

STORM :
ln that case , why ?
— ln that case , why ?

But ... ( screaming )
But ...

Sasaki !
Dr. Sasaki !

My goodness .
Oh , my goodness .
```
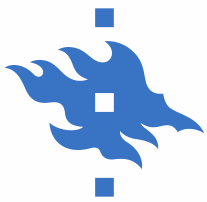
Alternative Subtitles

" I would like to receive an answer . "
" I would like to have an answer . "

" As mother wants it . "
" As mother wishes . "

" I must say everything ...
" I must tell everything ...

" By marrying me , you wanted to smother all this . "
" By marrying me , you wanted to hide all this . "

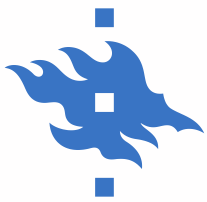And he went from village to village , feeding on charity .
And he went from village to village , living off alms .

" Let me have her to myself now that she is dying . "
Now that she 's dying , let me have her to myself .

" l shall go fetch him at once . "
I shall go find him at once . "

Alternative Subtitles

" We don 't have the sterilising oven going yet . "
It may be full of germs , and our sterilizing oven isn 't working yet .

" Did you mend my coat ? "
Was it you who mended my coat ' ? "

" Captive , take leave of thy prison ! "
" Captive , come forth from thy prison !

" For a year thou wilt reap souls . "
For a year thou wilt set souls free from the earthly realm .

" lt is Death , but too soon . "
It is Death , but he has come too soon . "

What 's the matter ?
What 's wrong ?

I 'm gonna fire some of those people .
I 'm gonna fire some of them .

Internal XML Format

```xml
<?xml version="1.0" encoding="utf-8"?>
<document>
  <CHAPTER ID="0">
    <P id="1"></P>
    <SPEAKER ID="1" LANGUAGE="DE" NAME="Rübig">
      <P id="2">
        <s id="1">Madam President, I saw a few boats landing at
                   Parliament this week and notified the security
                   service.</s>
        <s id="2">Not only were there language difficulties; the
                   telephone line was so poor that it was almost
                   impossible to communicate.</s>
        <s id="3">I would be most obliged if the number on which
                   the security service can be reached could also
                   be clearly displayed in the House, so that if
                   anyone wants to report an incident, they can do
                   so quickly and efficiently.</s>
      </P>
      <P id="3"></P>
    </SPEAKER>
    ...
```

Sentence Alignment

```xml
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">
<cesAlign version="1.0">

  <linkGrp targType="s" fromDoc="en/ep-00-01-17.xml.gz"
                        toDoc="fr/ep-00-01-17.xml.gz">

    <link xtargets="1;1" />
    <link xtargets="2;2" />
    <link xtargets="3;3 4" />
    ...
```

aligns sentence 3 with
sentences 3 and 4

Sentence Alignment

```xml
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">
<cesAlign version="1.0">

 <linkGrp targType="s" toDoc="fr/2005/CES_AC71_2005_5SUMMARY.xml.gz"
                       fromDoc="en/2005/CES_AC71_2005_5SUMMARY.xml.gz">

   <link certainty="0"        xtargets=";1 2" id="SL1" />
   <link certainty="0.612088" xtargets="1;3"  id="SL2" />
   <link certainty="0.173077" xtargets="2;4"  id="SL3" />
   <link certainty="1.65065"  xtargets="3;5"  id="SL4" />
   <link certainty="1.63824"  xtargets="4;6"  id="SL5" />
   <link certainty="-0.3"     xtargets=";7"   id="SL6" />
```

alignment scores

insertion of a target language sentence

Linguistic Markup

```xml
<?xml version="1.0" encoding="utf-8"?>
<document><CHAPTER ID="0">
<P id="1" /><SPEAKER ID="1" LANGUAGE="DE" NAME="Rübig"><P id="2">
<s id="1">
 <chunk type="NP" id="c-1">
  <w hun="NNP" tree="NN" lem="madam" pos="NNP" id="w1.1">Madam</w>
  <w hun="NNP" tree="NP" lem="President" pos="NNP" id="w1.2">President</w>
 </chunk>
 <w hun="," tree="," lem="," pos="," id="w1.3">,</w>
 <chunk type="NP" id="c-3">
  <w hun="PRP" tree="PP" lem="I" pos="PRP" id="w1.4">I</w>
 </chunk>
 <chunk type="VP" id="c-4">
  <w hun="VBD" tree="VVD" lem="see" pos="VBD" id="w1.5">saw</w>
 </chunk>
 ...
```
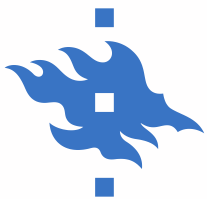
TMX Files

```xml
<tmx version="1.4">
<header creationdate="Fri Aug 23 10:17:33 2013"
          srclang="en"
          adminlang="en"
          o-tmf="unknown"
          segtype="sentence"
          creationtool="Uplug"
          creationtoolversion="unknown"
          datatype="PlainText" />
  <body>
    <tu>
      <tuv xml:lang="en"><seg>Statement of Government Policy by the Prime Minister, Mr Ingvar Carlsson, at the Opening of the Swedish Parliament on Tuesday, 4 October, 1988.</seg></tuv>
      <tuv xml:lang="fr"><seg>Declaration de Politique Générale du Gouvernement présentée mardi 4 octobre 1988 devant le Riksdag par Monsieur Ingvar Carlsson, Premier Ministre.</seg></tuv>
    </tu>
    <tu>
      <tuv xml:lang="en"><seg>Your Majesties, Your Royal Highnesses, Mr Speaker, Members of the Swedish Parliament.</seg></tuv>
      <tuv xml:lang="fr"><seg>Majestés, Altesses Royales, Monsieur le Président, Mesdames et Messieurs les députés!</seg></tuv>
    </tu>
...
```

Moses Plain Text Format

```
RF.en-fr.en
RF.en-fr.fr
```

The contents of the English file looks like this:

```
Statement of Government Policy by the Prime Minister, Mr Ingvar Carlsson, at the Opening of
the Swedish Parliament on Tuesday, 4 October, 1988.
Your Majesties, Your Royal Highnesses, Mr Speaker, Members of the Swedish Parliament.
Sweden's policy of neutrality is of decisive importance for our peace and independence.
It also contributes to stability and détente in our part of the world.
There is wide popular support for this policy.
It will be pursued with firmness and consistency.
...
```

And the corresponding French file looks like this:

```
Declaration de Politique Générale du Gouvernement présentée mardi 4 octobre 1988 devant le
Riksdag par Monsieur Ingvar Carlsson, Premier Ministre.
Majestés, Altesses Royales, Monsieur le Président, Mesdames et Messieurs les députés!
La politique suédoise de neutralité revêt une importance capitale pour la paix et l'
indépendance de notre pays.
Elle contribue également à la stabilité et à la détente dans notre secteur du monde.
Cette politique recueille une large adhésion populaire.
Elle sera poursuivie avec énergie et cohérence.
...
```

What do you think?