OPUS

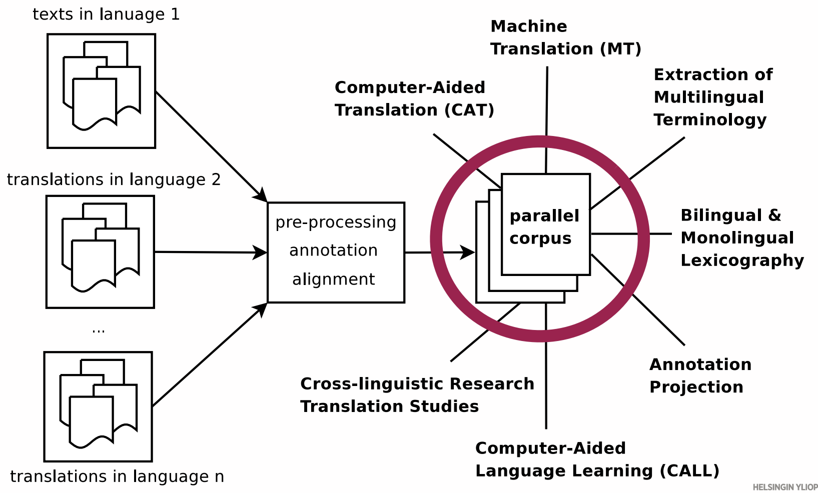
Parallel Corpora for Everyone

(NLPL-Activity G)

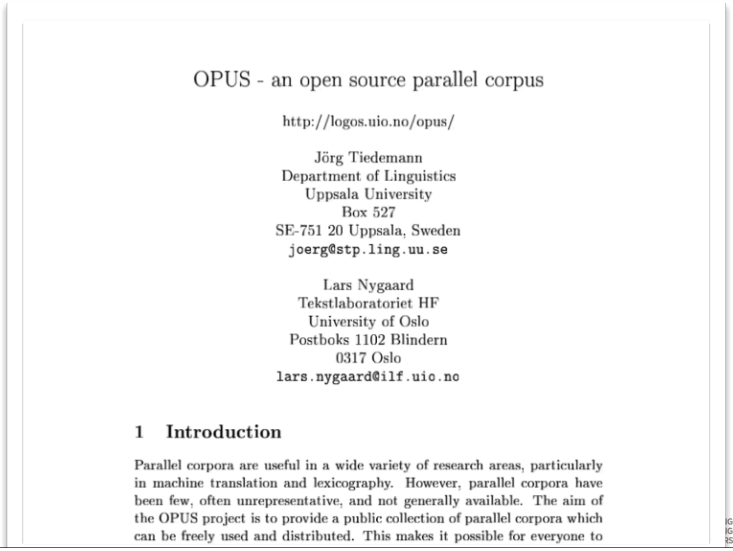
Jörg Tiedemann
Department of Digital Humanities
University of Helsinki

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
Computer-Aided Language Learning (CALL)
Cross-linguistic Research Translation Studies

NoDaLiDa 2003 in Reykjavík



OPUS - an open source parallel corpus

http://logos.uio.no/opus/

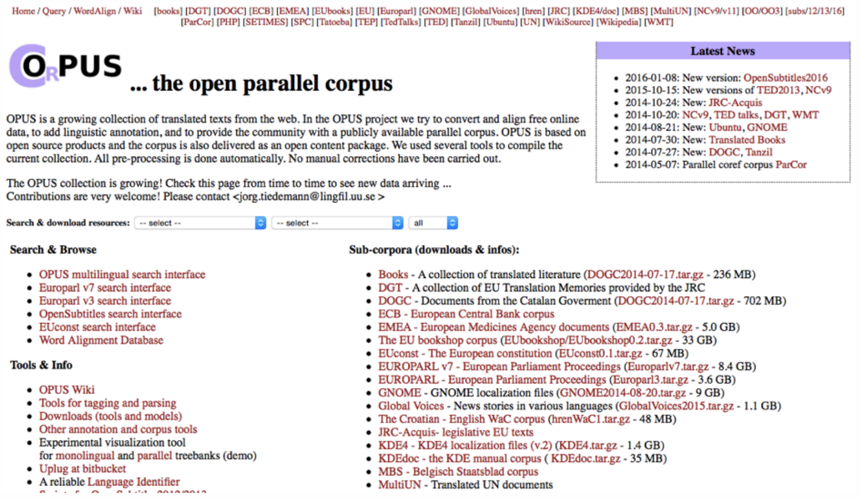
Jörg Tiedemann
Department of Linguistics
Uppsala University
Box 527
SE-751 20 Uppsala, Sweden
joerg@stp.ling.uu.se

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.nlpl.eu



Home / Query / WordAlign / Wiki [books] [DGT] [DOGC] [ECB] [EMEA] [EUbooks] [EU] [Europarl] [GNOME] [GlobalVoices] [hren] [JRC] [KDE4/doc] [MBS] [MultiUN] [NCv9/v11] [OO/OO3] [subs/12/13/16] [ParCor] [PHP] [SETIMES] [SPC] [Tatoeba] [TEP] [TedTalks] [TED] [Tanzil] [Ubuntu] [UN] [WikiSource] [Wikipedia] [WMT]

OPUS ... 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

OPUS

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

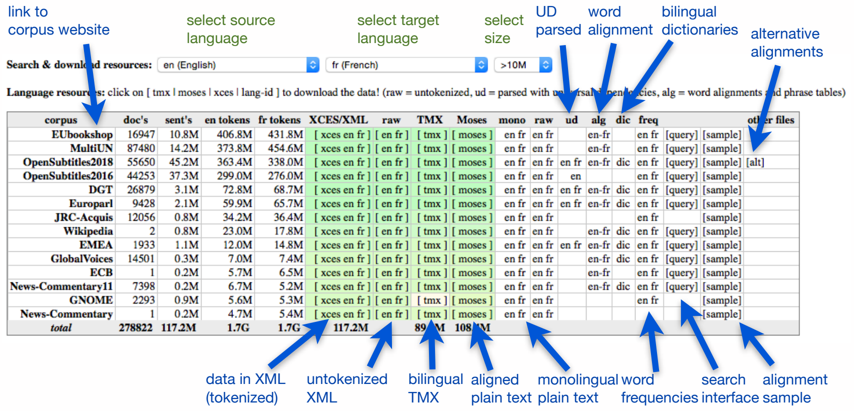
- tools for conversion, annotation, alignment
- online search interfaces
- word alignment database + concordances

http://opus.nlpl.eu

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

How To Find Resources

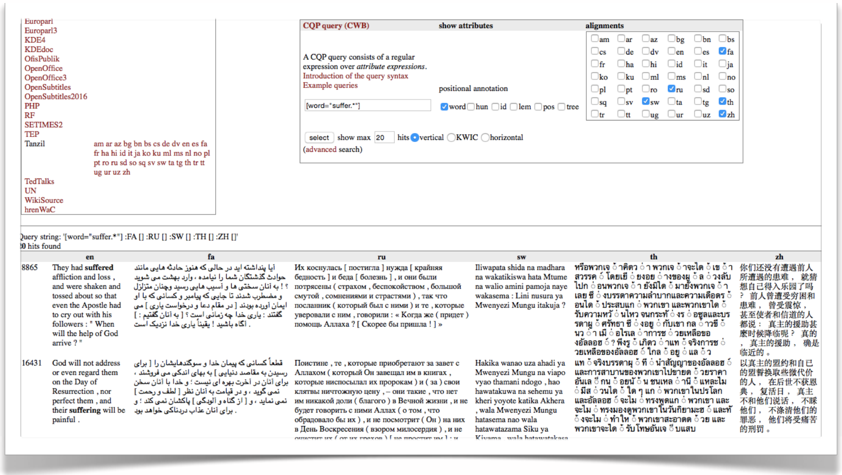


Search & download resources: en (English) | fr (French) | >10M

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

corpus	doc's	sent's	en tokens	fr tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq			other files
EUbookshop	16947	10.8M	406.8M	431.8M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr		en-fr		en fr	[query]	[sample]	
MultiUN	87480	14.2M	373.8M	454.6M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr		en-fr		en fr	[query]	[sample]	
OpenSubtitles2018	55650	45.2M	363.4M	338.0M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr	en fr	en-fr	dic	en fr	[query]	[sample]	[alt]
OpenSubtitles2016	44253	37.3M	299.0M	276.0M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr	en			en fr	[query]	[sample]	
DGT	26879	3.1M	72.8M	68.7M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr	en fr	en-fr	dic	en fr	[query]	[sample]	
Europarl	9428	2.1M	59.9M	65.7M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr	en fr	en-fr	dic	en fr	[query]	[sample]	
JRC-Acquis	12056	0.8M	34.2M	36.4M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr				en fr		[sample]	
Wikipedia	2	0.8M	23.0M	17.8M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr		en-fr	dic	en fr	[query]	[sample]	
EMEA	1933	1.1M	12.0M	14.8M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr	en fr	en-fr	dic	en fr	[query]	[sample]	
GlobalVoices	14501	0.3M	7.0M	7.4M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr		en-fr	dic	en fr	[query]	[sample]	
ECB	1	0.2M	5.7M	6.5M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr		en-fr		en fr	[query]	[sample]	
News-Commentary11	7398	0.2M	6.7M	5.2M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr		en-fr	dic	en fr	[query]	[sample]	
GNOME	2293	0.9M	5.6M	5.3M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr				en fr		[sample]	
News-Commentary	1	0.2M	4.7M	5.4M	[xces en fr]	[en fr]	[tmx]	[moses]	en fr	en fr						[sample]	
total	278822	117.2M	1.7G	1.7G	117.2M		89[illegible]M	108[illegible]M									

On-Line Search



Europarl
Europarl3
KDE4
KDEdoc
OfisPublik
OpenOffice
OpenOffice3
OpenSubtitles
OpenSubtitles2016
PHP
RF
SETIMES2
TEP
Tanzil — am ar az bg bn bs cs de dv en es fa fr ha hi id it ja ko ku ml ms nl no pl pt ro ru sd so sq sv sw ta tg th tr tt ug ur uz zh
TedTalks
UN
WikiSource
hrenWaC

CQP query (CWB) — show attributes — alignments

A CQP query consists of a regular expression over *attribute expressions*.
Introduction of the query syntax
Example queries

`[word="suffer.*"]`

positional annotation: ☑ word ☐ hun ☐ id ☐ lem ☐ pos ☐ tree

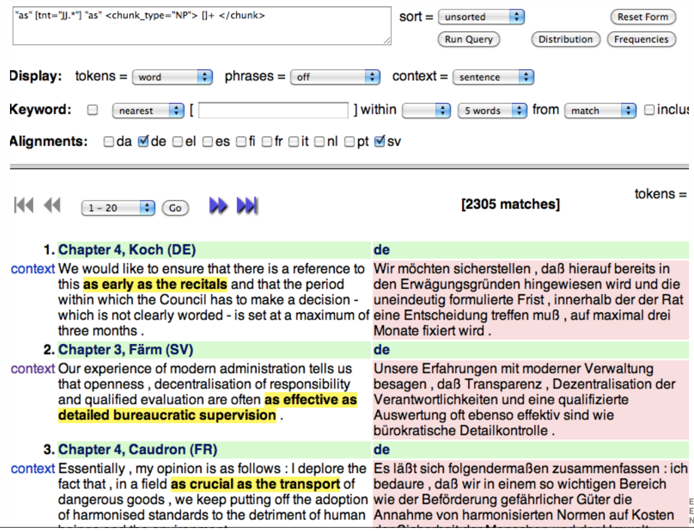
alignments: ☐ am ☐ ar ☐ az ☐ bg ☐ bn ☐ bs ☐ cs ☐ de ☐ dv ☐ en ☐ es ☑ fa ☐ fr ☐ ha ☐ hi ☐ id ☐ it ☐ ja ☐ ko ☐ ku ☐ ml ☐ ms ☐ nl ☐ no ☐ pl ☐ pt ☐ ro ☑ ru ☐ sd ☐ so ☐ sq ☐ sv ☑ sw ☐ ta ☐ tg ☑ th ☐ tr ☐ tt ☐ ug ☐ ur ☐ uz ☑ zh

select show max 20 hits ● vertical ○ KWIC ○ horizontal
(advanced search)

Query string: '[word="suffer.*"] :FA [] :RU [] :SW [] :TH [] :ZH []'
20 hits found

	en	fa	ru	sw	th	zh
8865	They had **suffered** affliction and loss , and were shaken and tossed about so that even the Apostle had to cry out with his followers : " When will the help of God arrive ? "	آیا پنداشته اید در حالی که هنوز حادثه هایی مانند حوادث گذشتگان شما را نیامده ، وارد بهشت می شوید ؟ ! به آنان سختی ها و آسیب هایی رسید وچنان متزلزل و مضطرب شدند تا جایی که پیامبر و کسانی که با او ایمان آورده بودند [در مقام دعا و درخواست یاری] می گفتند : یاری خدا چه زمانی است ؟ [به آنان گفتیم :] آگاه باشید ! یقیناً یاری خدا نزدیک است .	Их коснулась [постигла] нужда [крайняя бедность] и беда [болезнь] , и они были потрясены (страхом , беспокойством , большой смутой , сомнениями и страстями) , так что посланник (который был с ними) и те , которые уверовали с ним , говорили : « Когда же (придет) помощь Аллаха ? [Скорее бы пришла !] »	Iliwapata shida na madhara na wakatikiswa hata Mtume na walio amini pamoja naye wakasema : Lini nusura ya Mwenyezi Mungu itakuja ?	หรือพวกเจ ้าคิดว ่า พวกเจ ้าจะได ้เข ้าสวรรค ์ โดยเยี ่ยงอย ่างของผู ้ล ่วงลับไปก ่อนพวกเจ ้า ยังมิได ้มายังพวกเจ ้าเลย ซึ ่งบรรดาความลำบากและความเดือดร ้อนได ้ประสบแก ่พวกเขา และพวกเขาได ้รับความหวั ่นไหว จนกระทั ่งร ่อซูลและบรรดาผู ้ศรัทธา ซึ ่งอยู ่กับเขา กล ่าวว ่า เมื ่อไรเล ่าการช ่วยเหลือของอัลลอฮ ์? พึงรู ้เถิด ว ่าแท ้จริงการช ่วยเหลือของอัลลอฮ ์ ใกล ้อยู ่แล ้ว	你们还没有遭遇前人所遭遇的患难，就猜想自己得入乐园了吗？前人曾遭受穷困和患难，曾受震惊，甚至使者和信道的人都说：真主的援助甚麽时候降临呢？真的，真主的援助，确是临近的。
16431	God will not address or even regard them on the Day of Resurrection , nor perfect them , and their **suffering** will be painful .	قطعاً کسانی که پیمان خدا و سوگندهایشان را [برای رسیدن به مقاصد دنیایی] به بهای اندکی می فروشند ، برای آنان در آخرت بهره ای نیست ؛ و خدا با آنان سخن نمی گوید ، و در قیامت به آنان نظر [لطف و رحمت] نمی نماید ، و [از گناه و آلودگی] پاکشان نمی کند ؛ و برای آنان عذاب دردناکی خواهد بود .	Поистине , те , которые приобретают за завет с Аллахом (который Он завещал им в книгах , которые ниспосылал их пророкам) и (за) свои клятвы ничтожную цену , – они такие , что нет им никакой доли (благого) в Вечной жизни , и не будет говорить с ними Аллах (о том , что обрадовало бы их) , и не посмотрит (Он) на них в День Воскресения (взором милосердия) , и не очистит их (от их грехов) [не простит им] ; и	Hakika wanao uza ahadi ya Mwenyezi Mungu na viapo vyao thamani ndogo , hao hawatakuwa na sehemu ya kheri yoyote katika Akhera , wala Mwenyezi Mungu hatasema nao wala hatawatazama Siku ya Kiyama , wala hatawatakasa	แท ้จริงบรรดาผู ้ที ่นำสัญญาของอัลลอฮ ์และการสาบานของพวกเขาไปขายด ้วยราคาอันเล็กน ้อยนั ้น ชนเหล ่านี ้แหละไม ่มีส ่วนได ้ใด ๆ แก ่พวกเขาในปรโลก และอัลลอฮ ์จะไม ่ทรงพูดแก ่พวกเขา และจะไม ่ทรงมองดูพวกเขาในวันกิยามะฮ ์ และทั ้งจะไม ่ทำให ้พวกเขาสะอาดด ้วย และพวกเขาจะได ้รับโทษอันเจ็บแสบ	以真主的盟约和自己的盟誓换取些微代价的人，在后世不获恩典，复活日，真主不和他们说话，不睬他们，不涤清他们的罪恶，他们将受痛苦的刑罚。

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

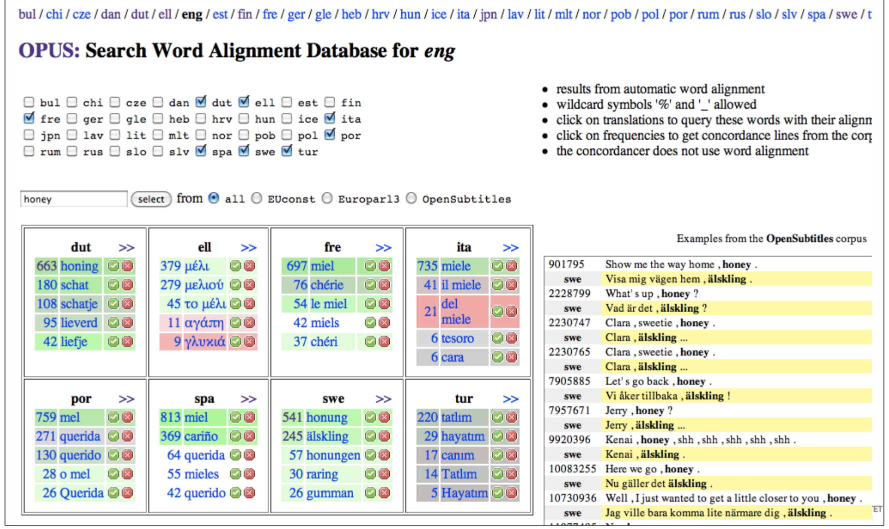
Keyword: ☐ nearest [ ] within 5 words from match ☐ inclu

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

1 - 20 Go — **[2305 matches]** — tokens =

		de
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		ell		fre		ita	
663	honing	379	μέλι	697	miel	735	miele
180	schat	279	μελιού	76	chérie	41	il miele
108	schatje	45	το μέλι	54	le miel	21	del miele
95	lieverd	11	αγάπη	42	miels	6	tesoro
42	liefje	9	γλυκιά	37	chéri	6	cara

por		spa		swe		tur	
759	mel	813	miel	541	honung	220	tatlım
271	querida	369	cariño	245	älskling	29	hayatım
130	querido	64	querida	57	honungen	17	canım
28	o mel	55	mieles	30	raring	14	Tatlım
26	Querida	42	querido	26	gumman	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** .

File Structure in OPUS

On Taito: **/proj/nlpl/data/OPUS/corpus**

```
Europarl
Europarl/xml
Europarl/xml/en
Europarl/xml/en/ep-10-07-05-004.xml.gz
Europarl/xml/fr
Europarl/xml/fr/ep-10-07-05-004.xml.gz
...
Europarl/xml/en-fr.xml.gz
...
Europarl/raw
Europarl/raw/en
Europarl/raw/en/ep-10-07-05-004.xml.gz
Europarl/raw/fr
Europarl/raw/fr/ep-10-07-05-004.xml.gz
...
Europarl/parsed
```

Internal XML Format

```
<?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 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 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

UD-Parsed Corpora (UDPipe)

```
<?xml version="1.0" encoding="utf-8"?>

<document>
  <CHAPTER ID="002">
    <P id="1">
      <s id="1">
        <w xpos="NOUN" head="1.2" feats="Number=Plur" upos="NOUN"
lemma="document" id="1.1" deprel="nsubj">Documents</w>
        <w xpos="VERB" head="0" feats="Mood=Ind|Tense=Past|
VerbForm=Fin" upos="VERB" misc="SpaceAfter=No" lemma="receive"
id="1.2" deprel="root">received</w>
        <w xpos="PUNCT" head="1.2" upos="PUNCT" lemma=":"
id="1.3" deprel="punct">:</w>
        <w xpos="VERB" head="1.2" feats="Mood=Imp|VerbForm=Fin"
upos="VERB" lemma="see" id="1.4" deprel="parataxis">see</w>
        <w xpos="PROPN" head="1.4" feats="Number=Plur"
upos="PROPN" misc="SpaceAfter=No" lemma="Minutes" id="1.5"
deprel="obj">Minutes</w>
      </s>
    </P>
  </CHAPTER>
</document>
```

Downloadable Packages

On Taito: **/proj/nlpl/data/OPUS/download**

```
Europarl/en.tar.gz
Europarl/en.raw.tar.gz
Europarl/en-pt.xml.gz
Europarl/en-pt.txt.zip
Europarl/en-pt.tmx.gz
```

On Abel: **/projects/nlpl/data/OPUS/download**

```
Europarl/en.raw.tar.gz
Europarl/en-pt.xml.gz
```

Word Alignment and Phrase Tables

On Taito: **/proj/nlpl/data/OPUS/wordalign**

```
wordalign/Europarl/en-fr/model/aligned.grow.gz
wordalign/Europarl/en-fr/model/aligned.srctotgt.gz
wordalign/Europarl/en-fr/model/aligned.grow-diag-final-and.gz
wordalign/Europarl/en-fr/model/aligned.intersect.gz
wordalign/Europarl/en-fr/model/aligned.tgttosrc.gz
wordalign/Europarl/en-fr/model/phrase-table-filtered.gz
wordalign/Europarl/en-fr/c.true.fr.gz
wordalign/Europarl/en-fr/c.true.en.gz
wordalign/Europarl/en-fr/ids.gz
```

```
en/0/1089124/4995691.xml.gz     fr/0/1089124/4588599.xml.gz     1       1
en/0/1089124/4995691.xml.gz     fr/0/1089124/4588599.xml.gz     2 3     2
en/0/1089124/4995691.xml.gz     fr/0/1089124/4588599.xml.gz     5 6     3
en/0/1089124/4995691.xml.gz     fr/0/1089124/4588599.xml.gz     7 8 9   4
...
```

OPUS Tools

On Taito: **module load nlpl-opus**

```
opus-read corpusname/lang-pair
opus-read -d corpusname lang-pair
opus-read -d corpusname -s srclang -t trglang

# print alignments with alignment certainty > LinkThr=0
opus-read -c 0 align-file.xml

# alignments with max 2 source sentences and 3 target sent's
opus-read -S 2 -T 3 align-file.xml
```

Links

Main webpage: http://opus.nlpl.eu

Documentation: http://opus.nlpl.eu/trac (wiki)

Search interfaces:

- http://opus.nlpl.eu/bin/opuscqp.pl
- http://opus.nlpl.eu/cwb/Europarl7/frames-cqp.html
- http://opus.nlpl.eu/cwb/OpenSubtitles/frames-cqp.html

Word alignment dictionary: http://opus.lingfil.uu.se/lex.php

Location in NLPL: $NLPL-HOME/data/OPUS