Brug af data på LetsMT-platformen

Dorte H. Hansen og Sussi Olsen,
Center for Sprogteknologi, Københavns Universitet

18. April 2012
Center for Sprogteknologi
Københavns Universitet

Hvilke data findes i LetsMT! Nu?

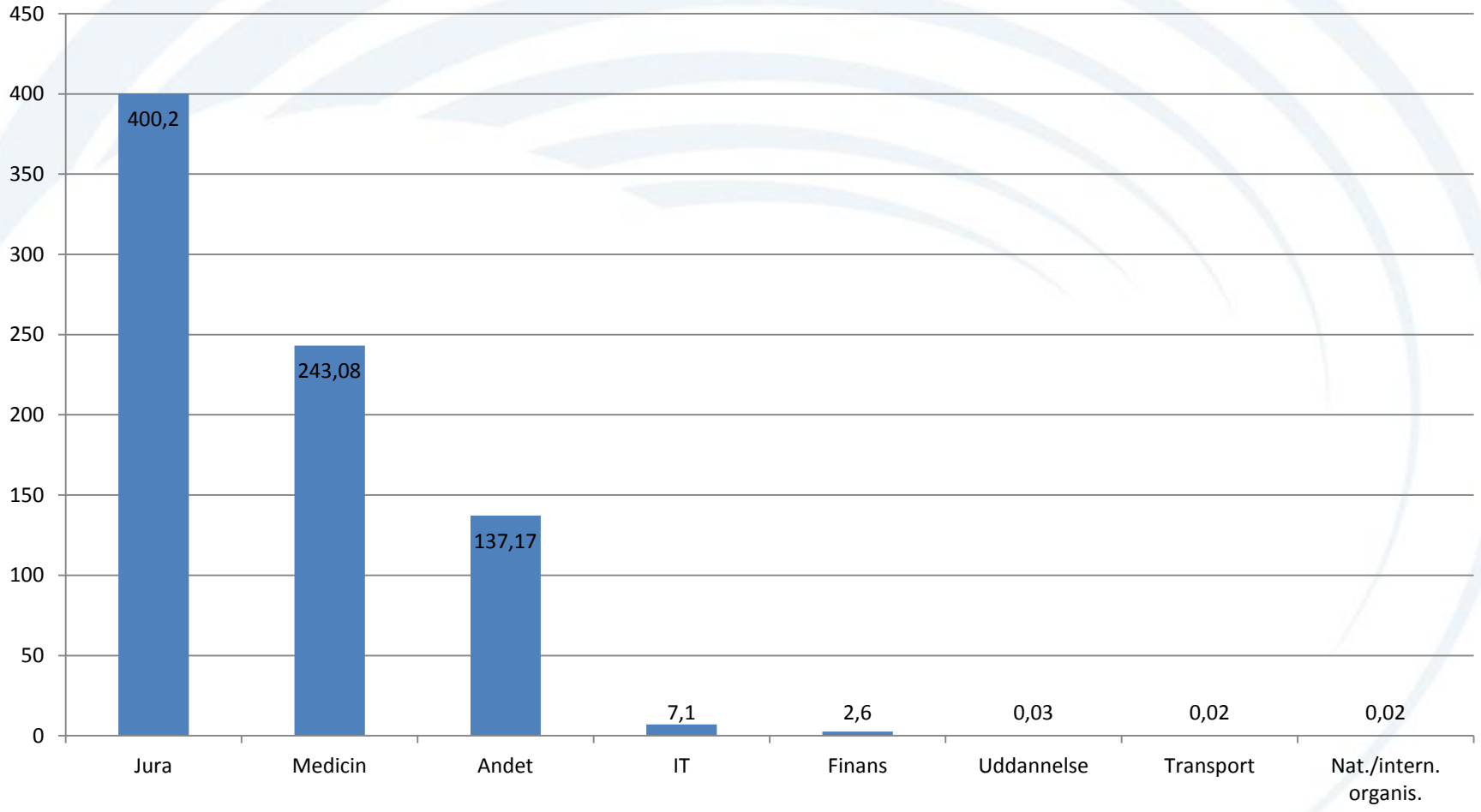
- Parallelle data:

  670 mio. parallelle sætninger (heraf 481 mio. offentlige) *

  Heraf en meget stor del EU-data
- 48 sprog (36 offentlige)
- 7 emneområder, af varierende størrelse:
  - jura
  - medicin og bioteknologi
  - finans
  - informationsteknologi og dataprocessering
  - nationale og internationale organisationer
  - (transport), kun private data
  - uddannelse
  - andet (også ved ubestemmeligt domæne eller blandede tekster)

* Tallene er ikke helt retvisende, samme sætninger tælles flere gange

Parallelle sætninger fordelt på antal domæner



Danske data på platformen

Let's MT! beta

Systems Corpora Translate Tools About Sign up Login

Corpora

This is a list of public and private corpora. Sign up to upload your own corpora and create translation systems.

Subject Domain - select - | Languages Danish | - add filter -

Name / Title	Subject Domain	Description	Size	Permissions
Semlab Business News 1	Finance	Business and finance news (annual reports) provided by Semlab	1.6M	Public
Cubes and Cones	Other	A small demo corpus.	<1k	Public
European Constitution (OPUS)	Law	A parallel corpus collected from the European Constitution (21 languages). Imported from OPUS.	2.1M	Public
European Medicine Agency	Biotechnology and health	European Medicine Agency	243M	Public
Eval, en-da finance	Finance	Automatic extracted evaluation corpus from the training of the English- Danish Finance III system ;	1k	Public
Evaluation, en-da finance	Finance	Automatic extracted evaluation corpus from the training of the English- Danish Finance III system ;	1k	Public
JRC-Acquis (v.3.0)	Law	The Acquis Communautaire (AC) is the total body of European Union (EU) law applicable the the EU Member States. This collection of legislative text changes continuously and currently comprises selected texts written between the 1950s and 2006.	204.3M	Public
KDE manuals (OPUS)	Information technology and data processing	A parallel corpus of KDE manuals (24 languages). Imported from OPUS.	0.3M	Public
Rapid 1 da-en, de-da	Other	Rapid press releases (http://europa.eu/rapid/) 5330 Danish-English and German Danish documents from 1993 -2003, tokenised and aligned with the HunAligner	0.3M	Public
Rapid 2 da-en, de-da	Other	Rapid press releases (http://europa.eu/rapid/); 2200 Danish-English documents from 2004 -2011, tokenised and aligned with the HunAligner	0.2M	Public
Semlab Business News 2	Finance	Business and finance news (annual reports) provided by Semlab	1M	Public
Uni-adm-1	National and international organizations and affairs	Danish-English administrative documents from Danish Universities. The data is tokenised and aligned with the Hunaligner	7.4k	Public
Uni-adm-2	Education	Danish-English curricula from Danish Universities. The data is tokenised and aligned with the Hunaligner	39.3k	Public

Danske data på platformen

- 71 mio. parallelle sætninger med dansk som det ene sprog alle offentlige (24.000 private)
- De store korpusser alle EU-data
- Antal sprogpar med dansk: > 20

Mono	en	de	fr	nl	it	es	pt	sv	el
10,9	4,4	4,2	4,2	4,1	4,1	4,1	4	3,2	2,9

Antal parallelle sætninger med dansk som det ene sprog (angivet i mio.)

- Mange mindre sprog som tjekkisk, ungarsk, baltiske sprog etc.
- Danske data inden for 6 domæner (- transport)

Dine egne data

- Hvordan kommer man i gang?
  - Gå ind på www.letsmt.eu
  - Sign up: send en anmodning til andris.karpovs@tilde.lv som derefter sender dig login data.

- Når du er blevet oprettet som bruger, kan du:
  - uploade dine tekster
  - træne dit eget MT-system
  - bruge andres offentlige tekster til træning

Beskriv dine egne data - metadata

Let's MT!

Metadata beskriver samlingen af filer (korpusset).

- Korpusnavn
- Korpustype (ensproget el. parallelt)
- Beskrivelse
- Domæne (14 stk., fx Law, Finance, Tourism, Education + other)
- Teksttype (9 stk. fx Financial Documentation + other)
- Tilladelse (public, private)

Du kan **ikke downloade** andres data

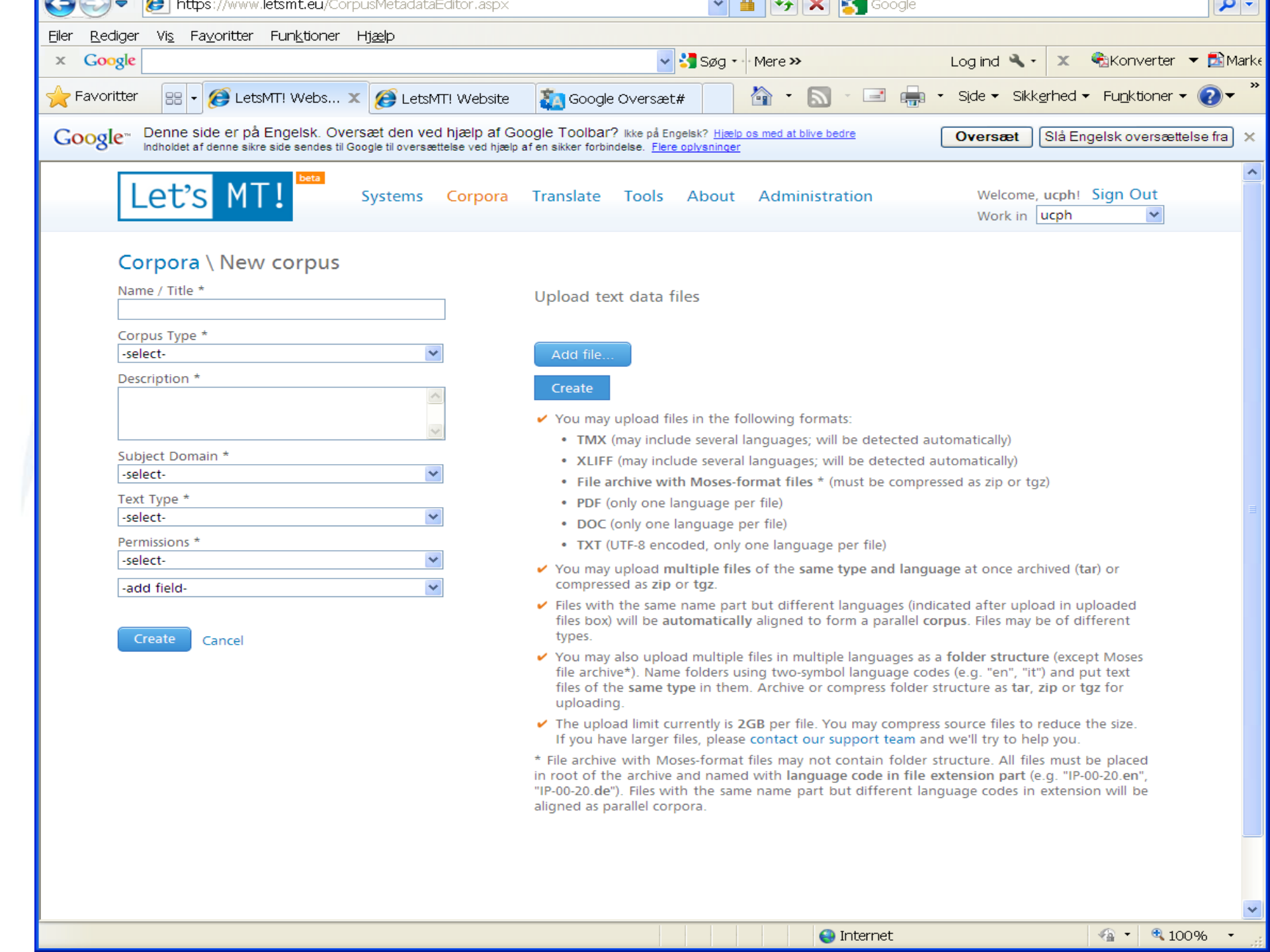
Du kan **kun se metadata** for andres data

Dine egne data III
- formater

LetsMT! Platformen understøtter:

- **TMX** (flersproget)
- **XLIFF** (flersproget)
- **Moses-format** (flersproget)
- **PDF** (ensproget)
- **DOC** (ensproget)
- **TXT** (ensproget)

De flersprogede formater kan uploades som zip-filer


https://www.letsmt.eu/CorpusMetadataEditor.aspx
Filer Rediger Vis Favoritter Funktioner Hjælp
Google
Søg
Mere »
Log ind
Konverter
Favoritter
LetsMT! Webs...
LetsMT! Website
Google Oversæt#
Side
Sikkerhed
Funktioner
Google™ Denne side er på Engelsk. Oversæt den ved hjælp af Google Toolbar? Ikke på Engelsk? Hjælp os med at blive bedre
Indholdet af denne sikre side sendes til Google til oversættelse ved hjælp af en sikker forbindelse. Flere oplysninger
Oversæt
Slå Engelsk oversættelse fra
Let's MT! beta
Systems
Corpora
Translate
Tools
About
Administration
Welcome, ucph! Sign Out
Work in ucph
Corpora \ New corpus
Name / Title *
Corpus Type *
-select-
Description *
Subject Domain *
-select-
Text Type *
-select-
Permissions *
-select-
-add field-
Create
Cancel
Upload text data files
Add file...
Create
You may upload files in the following formats:
TMX (may include several languages; will be detected automatically)
XLIFF (may include several languages; will be detected automatically)
File archive with Moses-format files * (must be compressed as zip or tgz)
PDF (only one language per file)
DOC (only one language per file)
TXT (UTF-8 encoded, only one language per file)
You may upload multiple files of the same type and language at once archived (tar) or compressed as zip or tgz.
Files with the same name part but different languages (indicated after upload in uploaded files box) will be automatically aligned to form a parallel corpus. Files may be of different types.
You may also upload multiple files in multiple languages as a folder structure (except Moses file archive*). Name folders using two-symbol language codes (e.g. "en", "it") and put text files of the same type in them. Archive or compress folder structure as tar, zip or tgz for uploading.
The upload limit currently is 2GB per file. You may compress source files to reduce the size. If you have larger files, please contact our support team and we'll try to help you.
* File archive with Moses-format files may not contain folder structure. All files must be placed in root of the archive and named with language code in file extension part (e.g. "IP-00-20.en", "IP-00-20.de"). Files with the same name part but different language codes in extension will be aligned as parallel corpora.
Internet
100%

Datakvalitet

God MT kræver træningsdata af god kvalitet

God kvalitet = direkte oversættelse:

DA	EN
Hos Vestas er ingen fejl for små til handling.	At Vestas, no error is too small to act on.

Dårlig kvalitet = forskellig sætnings- el. afsnitsrækkefølge (eller fri oversættelse) :

DA	EN
Vestas' rentedækningsgrad målt ved EBITDA var 15 i 2006.	As in 2006, Vestas will …
…	…
… garantihensatte i 2006 som planlagt.	**Vestas' interest cover ratio measured by EBITDA stood at 15 in 2006.**

Kvalitetssikring

God alignering er grundlaget for et godt MT-system

- Online alignering
  - På platformen aligneres de ensprogede filer (pdf-, doc- og txt-filer)
  - Lange sætninger frasorteres
  - Frasortering af ”0-alignments” (dvs. sætninger uden oversættelse)
- Offline alignering
  - Sikring af direkte oversættelse (parallelitet)
  - Sikring af fx pdf-til-txt-konvertering er gået godt
  - Frasortering af filer med dårlig alignering
  - Frasortering af ”0-alignments”
- Din egen TM (tmx-fil) er det sikreste at uploade

Hvor meget data er nok?

- Ofte siges: 1 mio. parallelle sætninger

- Vores tests viser dog:
  - 113.897 domænespecifikke sætninger (årsrapporter)

  → 60,91 BLEU

  - 19.415 domænespecifikke sætninger (ca. 400.000 ord, KU-data)
    + 506.887 almensproglige sætninger

  → 56,31 BLEU

BLEU = automatisk genereret mål for oversættelseskvalitet:

0 < meget dårligt < 30 dårligt < 50 kan efterredigeres < 80 godt < 100

Forholdet mellem data og kvalitet

Jo bedre kvalitet træningsdataene har,

og jo mere lighed der er mellem træningsdata og de data der senere skal oversættes ,

=> jo bedre oversættelseskvalitet får man