



# Open Services and Resources for Translation

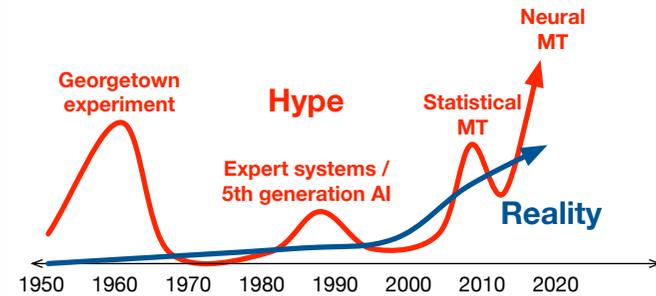
## The example of fiskmöö

Jörg Tiedemann

Language Technology, [jorg.tiedemann@helsinki.fi](mailto:jorg.tiedemann@helsinki.fi)  
 Department of Digital Humanities, University of Helsinki  
<https://blogs.helsinki.fi/language-technology/>



# The machine translation hype cycle



from Philipp Koehn: Neural Machine Translation



# ... but the practical side to it

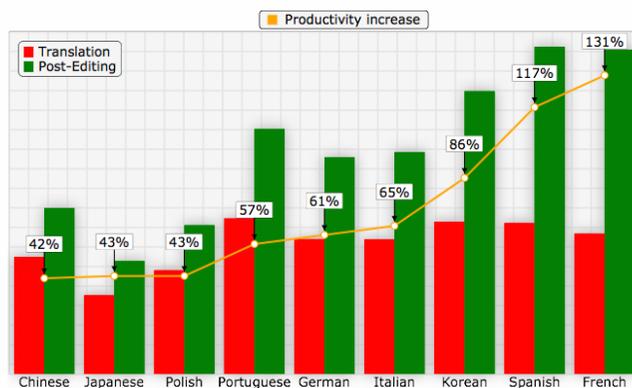
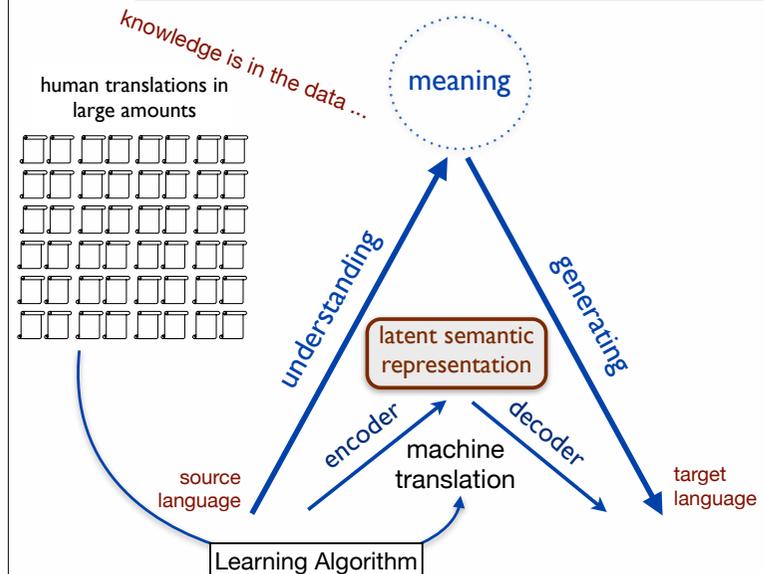


Figure 2.2: Increase in translator productivity when using machine translation (measured in words per hour). Study from Autodesk (Plitt and Masselot, 2010) on several language pairs with an in-house machine translation system.

from Philipp Koehn: Neural Machine Translation

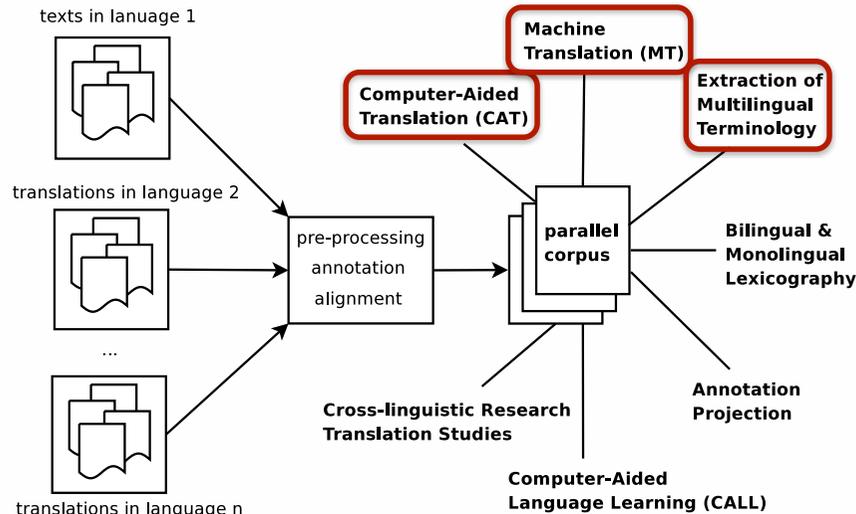


# Building machine translation





# The amazing utility of parallel corpora



# 18 years of MT resource development

- (1) Public data sets
- (2) Customizing machine translation
- (3) Free resources and translation models
- (4) New collaborations and use cases
- (5) Open translation services



# (1) OPUS - An Open Parallel corpus



NoDaLiDa 2003

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  
Tehtolaboratoriet HF  
University of Oslo  
Postboks 1102 Blindern  
0317 Oslo  
lars.nygaard@iif.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



<http://opus.nlpl.eu>

## Highlights

- over 800 language and language variants
- 5 billion sentences in 58 corpora
- several domains (legislation, medical, subtitles ...)
- > 76,000 language pairs, 10.8 billion translation units
- available in several data formats (OPUS XML, TMX, Moses)

## Tools & online services

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



# What is included?



- Copyright-free Books
- Bible translations
- 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
- JW300
- Belgisch Stadsblad
- United Nations corpora
- News Commentary, WMT sets
- OpenSubtitles
- ParaCrawl
- SETIMES
- Tatoeba
- TedTalks
- Tanzil Quran Translations
- Wikipedia, WikiSource
- ...



# Available resources

<http://opus.nlpl.eu>

link to corpus website    select source language    select target language    select size    UD parsed    word alignment    bilingual dictionaries    alternative alignments

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	[ query ]	[ sample ]					
MultiUN	87480	14.2M	373.8M	454.6M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
OpenSubtitles2018	55650	45.2M	363.4M	338.0M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]	[ alt ]				
OpenSubtitles2016	44253	37.3M	299.0M	276.0M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
DGT	26879	3.1M	72.8M	68.7M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
Europarl	9428	2.1M	59.9M	65.7M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
JRC-Acquis	12056	0.8M	34.2M	36.4M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
Wikipedia	2	0.8M	23.0M	17.8M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
EMEA	1933	1.1M	12.0M	14.8M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
GlobalVoices	14501	0.3M	7.0M	7.4M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
ECB	1	0.2M	5.7M	6.5M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
News-Commentary11	7398	0.2M	6.7M	5.2M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
GNOME	2293	0.9M	5.6M	5.3M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
News-Commentary	1	0.2M	4.7M	5.4M	[ xces en fr ]	[ en fr ]	[ tmx   moses ]	en fr	[ query ]	[ sample ]					
<b>total</b>	<b>278822</b>	<b>117.2M</b>	<b>1.7G</b>	<b>1.7G</b>	<b>117.2M</b>	<b>89 M</b>	<b>108M</b>								

data in XML (tokenized)    untokenized XML    bilingual TMX    aligned plain text    monolingual plain text    word frequencies    search interface    alignment sample

= training data for machine translation



# On-line search interfaces

<http://opus.nlpl.eu>

as [tree:"JJ\*" "as" <chunk\_type="NP"> [ + </chunk>

sort = [ unsorted ]    Run Query    Distribution    Frequencies

Display: tokens = [ word ]    phrases = [ off ]    context = [ sentence ]

Alignments: [ bg ] [ cs ] [ da ] [ de ] [ el ] [ es ] [ et ] [ fi ] [ fr ] [ hu ] [ it ] [ lv ] [ nl ] [ pl ] [ pt ] [ ro ] [ sk ] [ sl ] [ sv ]

1-20    Go    [3531 matches]

1. Chapter 4, Koch (DE)    fi    sv  
 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.    Me haluaisimme varmistaa, että tähän viitataan jo perusteluissa ja että moniselitteisestä tuotoitlu määräaika, jonka kuluessa neuvoston on tehtävä päätöksensä, määrätään enimmillään kolmeksi kuukaudeksi.    Vi skulle vilja säkerställa att det redan i motiveringen hänvisas till detta och att den icke entydigt formulerade tidsfristen, inom vilken rådet måste fatta beslut, fastställs till högst tre månader.

2. Chapter 4, Pieczyk (DE)    fi    sv  
 context On the subject at hand, I think that the people of Europe must be able to be confident that the goods - however dangerous they are - which are transported on Europe's roads, railways, and so on are as safe as possible.    Nyt asiaan: Euroopan kansalaisten on mielestäni voitava luottaa siihen, että se, mitä Euroopan maanteilla, rautateillä jne. kuljetetaan, jos kyse on vieläpä vaarallisista aineista, on mahdollisimman turvallista.    Till saken: Jag anser att Europas medborgare måste kunna lita på att det som transporteras på Europas vägar, järnvägsnät osv., även om det är farligt gods, transporteras så säkert som möjligt.

3. Chapter 3, FÄrrm (SV)    fi    sv  
 context Our experience of modern administration tells us that openness, decentralisation of responsibility and qualified evaluation are often as effective as detailed bureaucratic supervision.    Kokemuksemme nykyaikaisesta hallinnosta osoittavat, että avoimuus, vastuun hajauttaminen ja pätevä arviointi ovat usein yhtä tehokkaita kuin byrokraattinen yksityiskohtinen valvonta.    Våra erfarenheter av modern förvaltning säger oss att öppenhet, decentralisering av ansvaret och kvalificerad utvärdering ofta är lika effektiva som byråkratisk detaljkontroll.

4. Chapter 4, Caudron (FR)    fi    sv  
 directive on the transport of dangerous goods by road, dates from May 1999, however, and could not therefore take account of the latest comitology procedure.

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.

A further amendment allows the Member States to impose more stringent requirements, in particular for vacuum tanks, if work is done or goods are transported as a



# Multilingual lexical resources

bul / chi / cze / dan / dut / ell / eng / est / fin / fr / 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

- 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 corp
- 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 lieverde	11 γλυκύτη	42 miels	6 tesoro
42 liefje	9 γλυκύα	37 chéri	6 cara

Examples from the OpenSubtitles corpus

901795 Show me the way home ,honey .

2228799 Visa mig vägen hem , älskling .

2230747 Clara ,sweetie ,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 .

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 .

# Some external users of OPUS



ACL 2016-2019  
CONFERENCE ON  
MACHINE TRANSLATION (WMT)

Cognitive Analysis and Statistical Methods  
for Advanced Computer Aided Translation



SUBASUB

English → German ▶

Search for conversations in movies

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

# (2) The Let's MT project



<http://project.letsmt.eu>



<http://opus.npl.eu>

**Let's MT!**  
Platform for Online Sharing of Training Data and Building User Tailored Machine Translation

**META-NET**

**Main project objective**  
The core objective of Let'sMT! project is to provide innovative online MT services through sharing of parallel corpora provided by users, with emphasis on less-covered languages and specialized domains.

**Results**  
Let'sMT! has reached the following results:  
• Full version of the translation interface online  
• widget for free inclusion in a web page to provide its translation using Let'sMT! services  
• browser plugin for Internet Explorer and Mozilla Firefox for using Let'sMT! services  
• application for mobile devices using jQuery framework  
• integration of services in web pages  
• integration of services to computer-aided translation software

**Target users**  
The Let'sMT! project provides MT solutions for European citizens and businesses, allowing more efficient usage of multilingual corpora. The Let'sMT! platform targets actions in:  
• localization & translation industry - facilities for training MT systems on their data and generating custom MT systems to be used by localization providers or their clients;  
• users of business and financial news - free and instant MT services, with an emphasis on less-covered languages;  
• holders of language resources - easily building up MT services using their specific parallel data.

**Proposed services/solutions**  
The core of the Let'sMT! platform is the Moses SMT backends and GIZA++ . These are publicly available open source tools that are well known and widely used in statistical machine translation. The system also uses existing publicly available parallel corpora such as EU-Paralel, Europarl, OPUS and it will allow users to upload and build readily

**Technology**  
The core of the Let'sMT! platform is the Moses SMT backends and GIZA++ . These are publicly available open source tools that are well known and widely used in statistical machine translation. The system also uses existing publicly available parallel corpora such as EU-Paralel, Europarl, OPUS and it will allow users to upload and build readily

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

# MT data management system



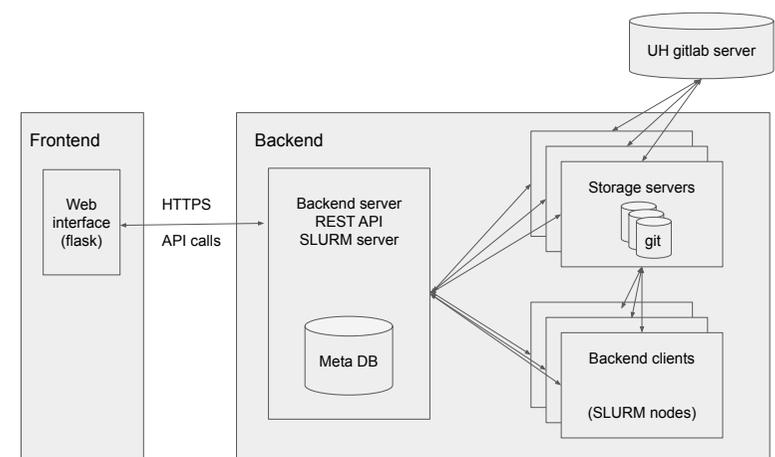
- data collection
- data extraction / conversion
- cross-lingual alignment
- data management



corpus repository

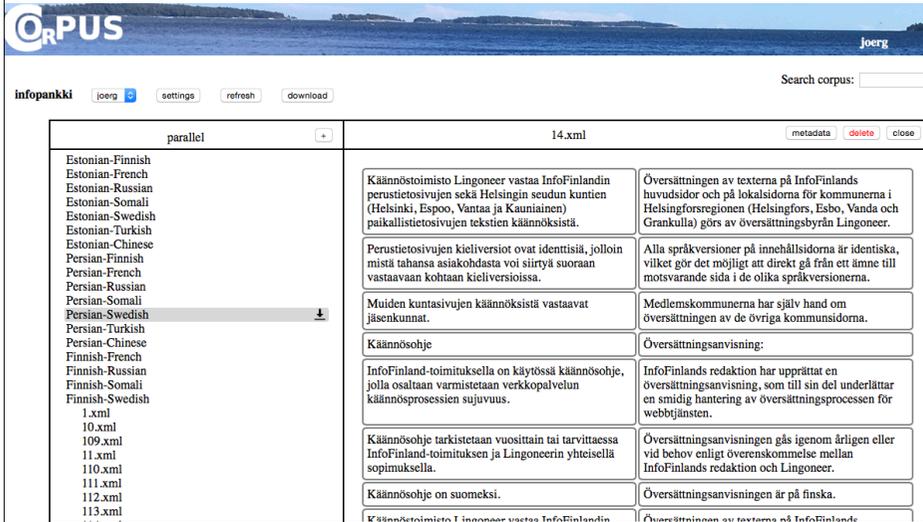
HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

# MT data management system





# MT data management system



<https://opus-repository.ling.helsinki.fi>

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI



## Tilde Custom Machine Translation

[Learn more →](#)

On the Tilde MT platform, you can translate files with your custom MT systems, add term collections to boost translation accuracy, and evaluate MT system quality – all in a single online workspace.

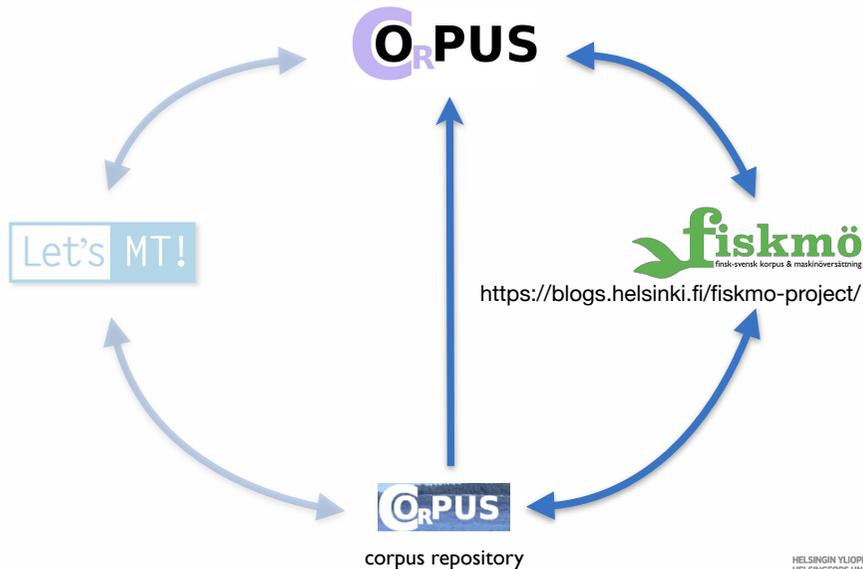
Don't have a custom MT system? Our MT team will develop a system for you, providing the full cycle of MT services. Or you can build a system yourself on the Tilde MT platform, using your data and resources from the Tilde Data Library.

Free 14-day trial

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI



## (3) The fiskmö project



HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI



### Our mission

- a **public** parallel Finnish-Swedish corpus for cross-lingual research with good coverage and high quality
- a **public** Finnish-Swedish translation service of high quality
- improved **tools for translators** in Finland to support public and private organizations to keep bilingual content alive





- alignment
- data curation
- machine translation
- user interfaces

Jörg Tiedemann   Mikko Aulamo



- web crawling
- bitext extraction

Filip Ginter   Jenna Kanerva



Akseli Leino



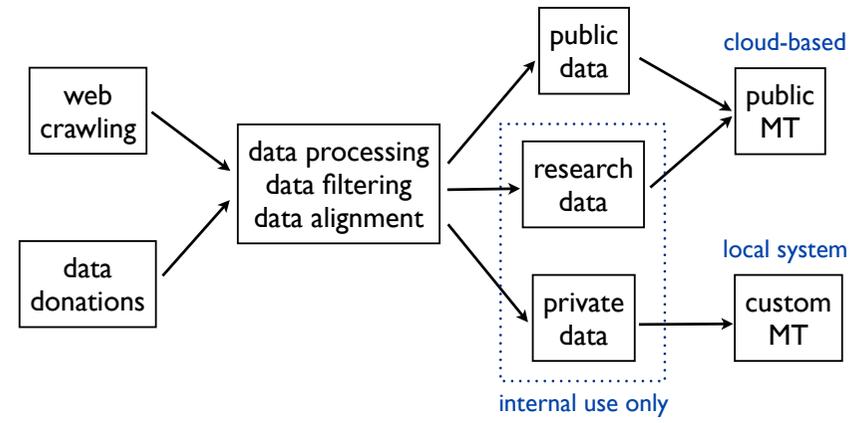
Tommi Nieminen   Niko Papula   Matias Tamminen



- user data collection
- workflow integration



### How does it work?



<https://blogs.helsinki.fi/fiskmo-project/>

## What are the benefits?

Why should you support us and what is in there for you?

- improve translation quality and quantity
- **clean your data** - we will help you
- we align your data and create clean translation memories
- get a **locally installable** customized translation engine with dedicated Trados plugin

This is non-profit, no costs, no commitment!  
Your data and translations are save and protected!



<https://blogs.helsinki.fi/fiskmo-project/>

The screenshot shows the fiskmö web interface. At the top, there are language selection buttons for Finnish, Swedish, Norwegian, and Danish. A 'detect language' button is also present. Below this is a 'translate' button with a right-pointing arrow. To the right, there are three upload options: 'upload translation memory (.tmx, .xliff)', 'upload translated documents (xml, html, txt, pdf, doc, srt, rtf, epub)', and 'upload translated web pages'. At the bottom left, there are input fields for 'original url:', 'translation url:', 'your email (optional):', and 'send TMX to your email:'. A red arrow points to the 'send TMX to your email:' field with the text 'send a translation memory'. At the bottom right, there are logos for Svenska kulturfonden, Helsingfors Universitet, and Language Technology. The URL <https://translate.ling.helsinki.fi> is displayed at the bottom right.



## Available data (translation memories)

<https://version.helsinki.fi/Helsinki-NLP/fiskmo>

Name	nr of TUs	availability	downloads
<a href="#">finlex/finlex-v2018.tmx</a>	1,311,996	public	z01, zip
<a href="#">infopankki/infopankki-v2019-04-05.tmx</a>	11,224	public	tmx
<a href="#">webcrawl/finnish-swedish-crawl-v1</a>	27,033	public	tmx
<a href="#">webcrawl/finnish-swedish-crawl-v2-clean</a>	500,000	public	tmx
<a href="#">webcrawl/finnish-swedish-crawl-v2-2M</a>	2,000,000	public	zip, z01
<a href="#">webcrawl/finnish-swedish-yle-rss-v2-clean</a>	25,000	public	tmx
<a href="#">webcrawl/finnish-swedish-yle-rss-v2-100K</a>	25,000	public	tmx
<a href="#">webcrawl/finnish-swedish-fiskmo-crawl-clean</a>	85,000	public	tmx
<a href="#">webcrawl/documents/finnish-swedish-fiskmo-crawl-articles-v1-0.8.tmx.zip</a>	199,242	public	tmx-small
<a href="#">webcrawl/documents/finnish-swedish-fiskmo-crawl-articles-v1-0.5.tmx.zip</a>	321,559	public	tmx-medium
<a href="#">webcrawl/documents/finnish-swedish-fiskmo-crawl-articles-v1.tmx.zip</a>	472,804	public	tmx-large
<a href="#">vnk/Budjettikatsaus.tmx</a>	1,115	public	tmx
<a href="#">vnk/Hallituksen_vuosikertomus.tmx</a>	14,288	public	tmx



## Embedded fiskmö MT in Trados Studio

The screenshot shows the Trados Studio interface with the fiskmö MT engine embedded. The main editor displays a Swedish text segment: 'Republiken president utnämnde Finlandens 75:e regering torsdagen den 6 juni.' The target text in Finnish is: 'tasavallan presidentti nimitti torstaina 6. kesäkuuta Suomen 75. hallituksen.' A 'Translation Results' window is open, showing the source and target text segments. The interface includes a menu bar with options like File, Home, Review, Advanced, View, Add-Ins, and Help. The status bar at the bottom shows '0,00%' and 'Chars: 77 / 0/801'.

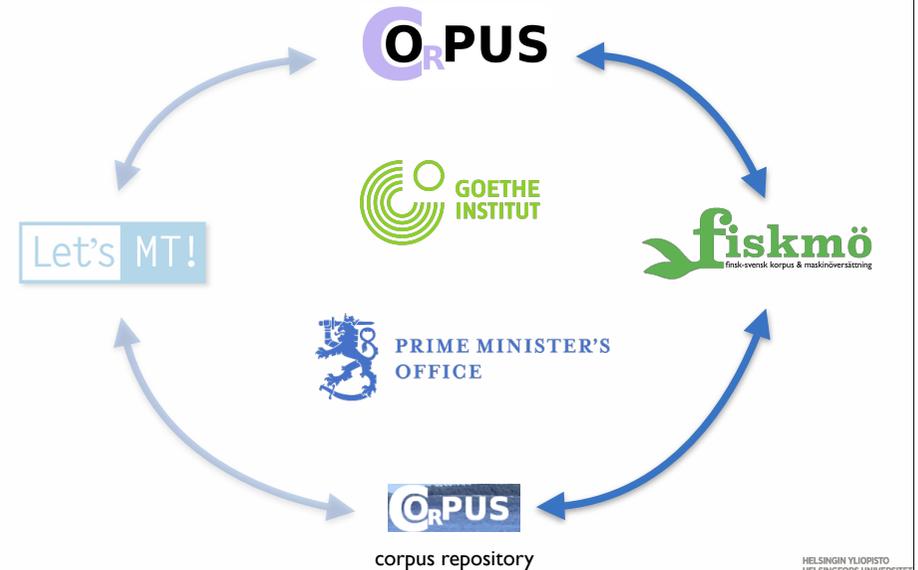
## How good is it now?

A recent result with our fiskmö benchmark test set:

- |                           |             |              |
|---------------------------|-------------|--------------|
| • from Finnish to Swedish | BLEU        | chr-F        |
| • Google Translate        | 17.7        | 0.513        |
| • Presidency MT           | <b>29.5</b> | <b>0.627</b> |
| • fiskmö                  | 26.6        | 0.600        |
| • from Swedish to Finnish | BLEU        | chr-F        |
| • Google Translate        | 19.2        | 0.564        |
| • Presidency MT           | 25.6        | 0.612        |
| • fiskmö                  | <b>26.1</b> | <b>0.623</b> |



## (4) Other Collaborations



## The collaboration with the Goethe Institut

### A practical tool for translating websites

- German to Finnish (and possibly other languages)
- tuned for the content on GI websites

### Language technology and AI

- Is machine translation capable of translating **creative text**?
- How does a machine cope with **racism** and **political correctness**?
- What kind of **bias** is in the model?



## A practical translation tool

### Data collection

- translated documents from the Goethe Institut
- aligning sentences with each other
- cleaning data

### Training machine translation

- state-of-the-art neural network models
- start with a general purpose model
- fine-tuning with GI data



<https://translate.ling.helsinki.fi/goethe>



German-Finnish (fine-tuned for GI) German-Finnish  
Finnish-German

tuned on only ca 6,700 sentence pairs

Die Nacht der Wissenschaften! Das wird ein phänomenales Ereignis!



Tieteen yö! Siitä tulee ilmiömäinen tapahtuma!

Support the project by providing additional training data:

upload translation memory (.tmx, .xliff)

upload translated documents (.xml, .html, .txt, .pdf, .doc, .srt, .rtf, .epub)

upload translated web pages



## The effect of customization

### Comparison to other systems

- reference-based evaluation (2000 sentences)

System	BLEU
Systran	12.1
Yandex	14.8
Google	17.6
Helsinki - no tuning	17.3
<b>Helsinki - tuned</b>	<b>21.4</b>



## Lessons learnt

### Data collection is hard (but very important!)

- only 10,000 sentence pairs in the end
- converting and cleaning takes time

### Translation performance

- we can easily beat general-purpose engines
- **idiomatic expressions** are difficult to translate
- missing **lexical coverage** (e.g. bird names ...)
- issues with **coherence**

## Outlook into the future

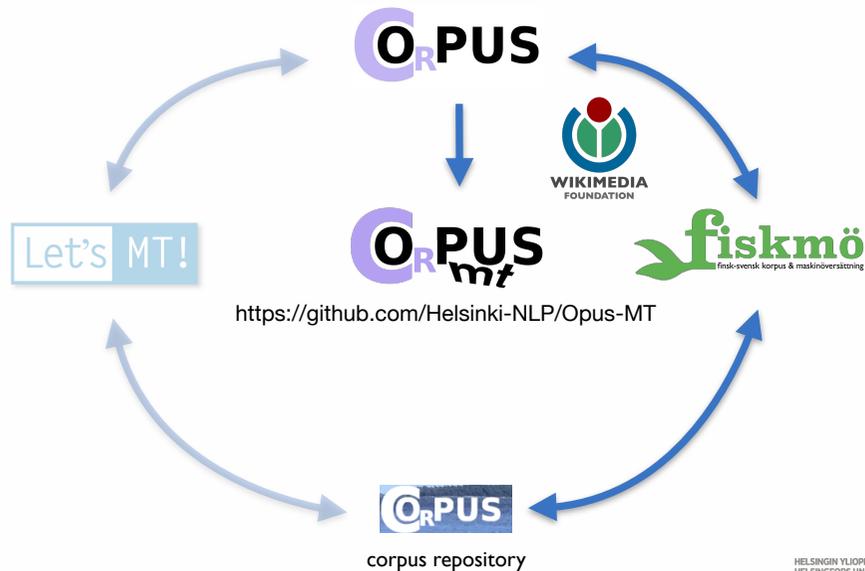
Roadmap to an open and free translation service

- free translation services for **many languages**
- **community effort** to build resources
- better coverage of **minority languages**
- distributed open environment

Equal access to information is a human right!



## (5) Open Translation Services



## OPUS-MT

<https://github.com/Helsinki-NLP/Opus-MT>



### Available software:

- MT server solution based on Marian NMT
- dockerized web-app
- simple on-line translation interface

### Pre-trained translation models:

- number of bilingual models: 1,042
- number of multilingual models: 43
- number of supported source languages: 180
- number of supported target languages: 173
- number of supported language pairs: 1,421

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI



## The future of OPUS-MT

<https://github.com/Helsinki-NLP/Opus-MT>



### More languages and more domains

- multilingual models and transfer learning
- low resource languages, dialects and language variants
- modular approach with exchangeable encoders/decoders
- simple fine-tuning and domain adaptation
- audio input / output

### Open translation services (wish-list)

- de-centralized user-contributed translation services
- automatic load balancing
- personalized services (with feedback loops)

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI



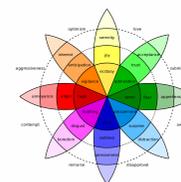
## Related research in Helsinki

Data collection for MT  
> 200 languages

<http://opus.nlpl.eu>

**FOTRAN**  
Found in Translation

learning semantics  
from > 1,000 languages



cross-lingual  
sentiment  
analysis



Data collection & MT  
2 languages (Finnish/Swedish)  
<https://blogs.helsinki.fi/fiskmo-project/>

audiovisual data & MT  
6 languages



**MeMAD**  
Methods for Managing  
Audiovisual Data  
<https://memad.eu>

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI



# Language Technology in Helsinki

<http://blogs.helsinki.fi/language-technology/>



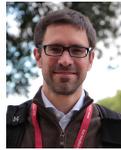
NLP for Finno-Ugric languages



FSTs  
Omorfi  
FinPos



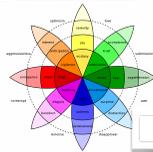
Language Technology for low-resource languages  
CALL & paraphrasing



Creative language generation



The Peace Machine



senitimentator



## LANGUAGE TECHNOLOGY

NLP Research at the University of Helsinki and Information about the Language T

NEWS RESEARCH EVENTS PEOPLE STUDY



## Language Technology in Helsinki

<http://blogs.helsinki.fi/language-technology/>

- Where are we? Metsätalo, Unioninkatu 40 B, Helsinki
- Software at <https://github.com/Helsinki-NLP>
- UH research portal: <https://researchportal.helsinki.fi/en/organisations/language-technology>
- Twitter: [@HelsinkiNLP](https://twitter.com/HelsinkiNLP)