Open Services and Resources for Translation

The example of fiskmö

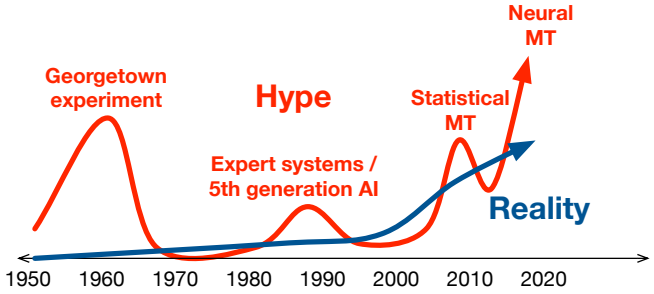
Jörg Tiedemann

Language Technology, 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

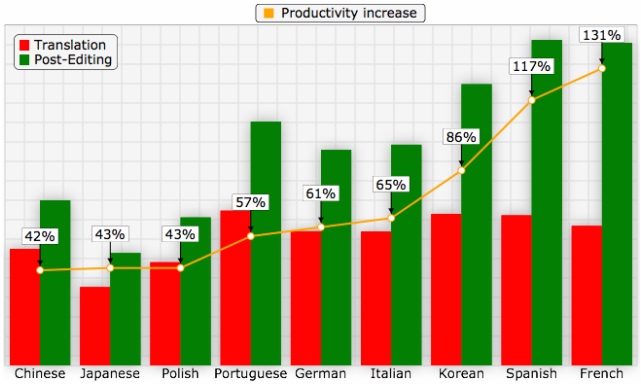
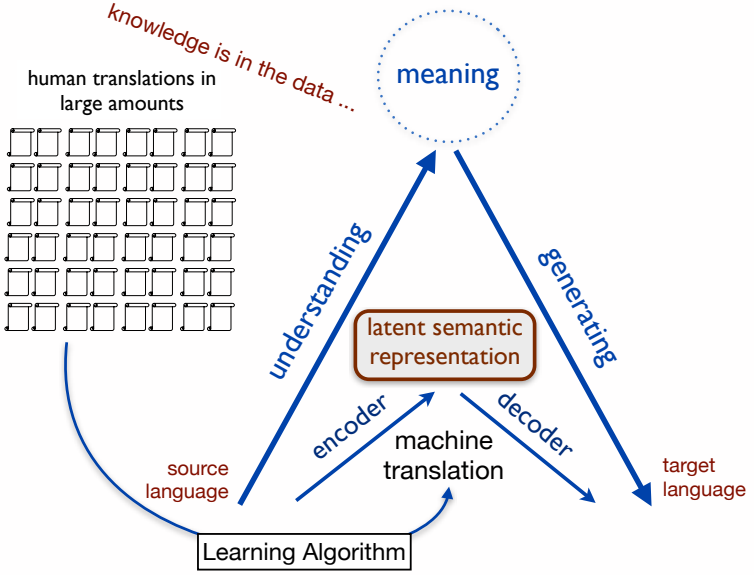


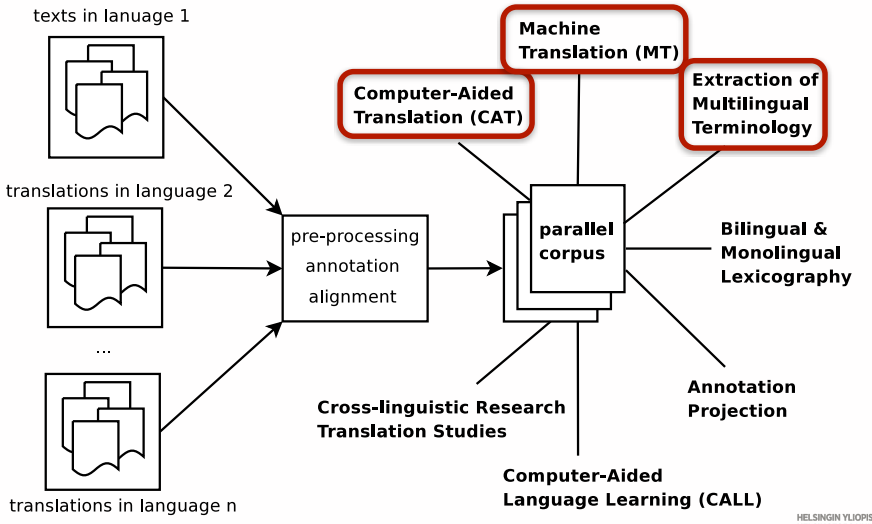
Figure 2.2: Increase in translator productivity when using machine translation (measured in words per hour). Study from Autodesk (Plitt and Masselot, 2010) on several language pairs with an in-house machine translation system.

from Philipp Koehn: Neural Machine Translation

Building machine translation



The amazing utility of parallel corpora



18 years of MT resource development

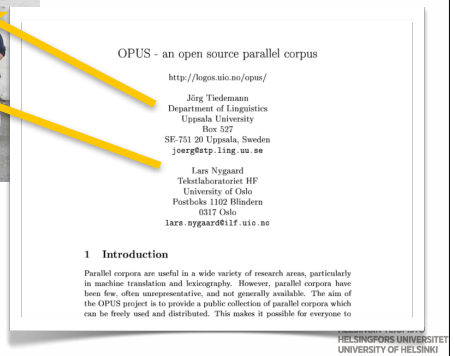
(1) Public data sets
(2) Customizing machine translation
(3) Free resources and translation models
(4) New collaborations and use cases
(5) Open translation services

(1) OPUS - An Open Parallel corpUS

OPUS
http://opus.nlpl.eu

Rosendal 2002

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
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

OPUS 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

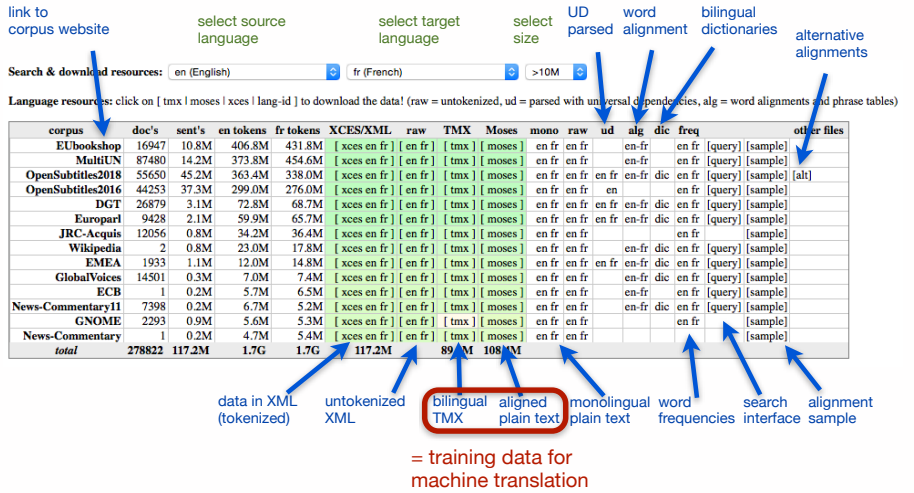
- tools for conversion, annotation, alignment
- online search interfaces
- word alignment database + concordances

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
- ...

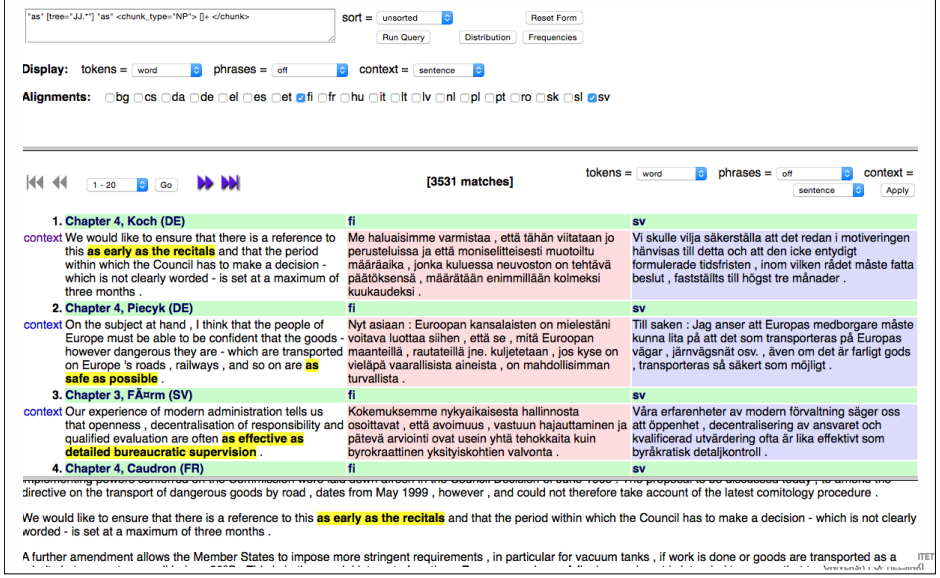
Available resources

http://opus.nlpl.eu

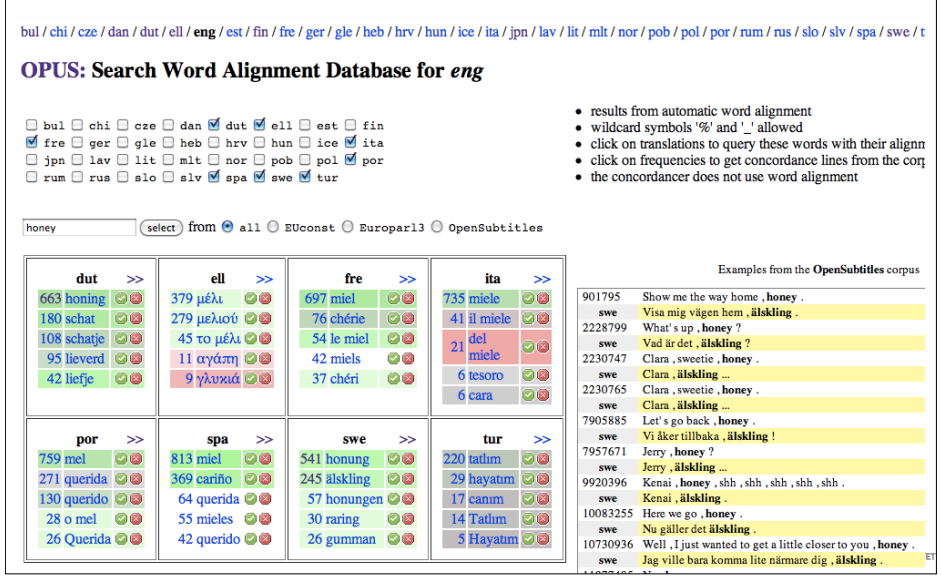


On-line search interfaces

http://opus.nlpl.eu



Multilingual lexical resources



Some external users of OPUS

CASMACAT

Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation

ACL 2016-2019
CONFERENCE ON
MACHINE TRANSLATION (WMT)

Let's MT!

Sketch Engine

Reverso Context
The best way to translate words and expressions

SUBaSUB

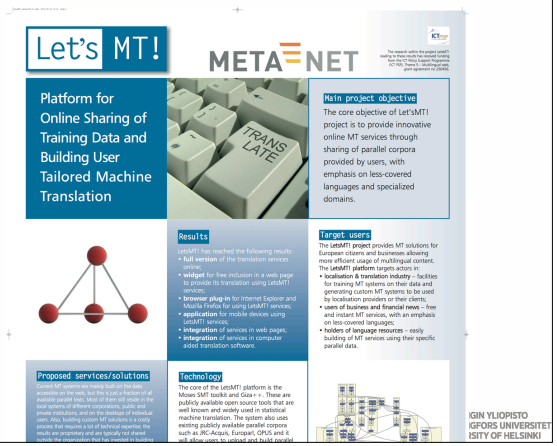
English↔German

Search for conversations in movies

(2) The Let's MT project

OPUS
http://opus.nlpl.eu

Let's MT!
http://project.letsmt.eu



MT data management system

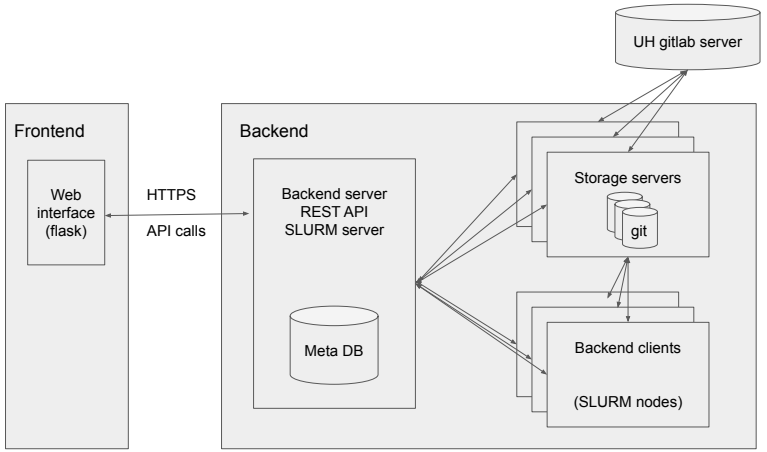
OPUS

Let's MT!

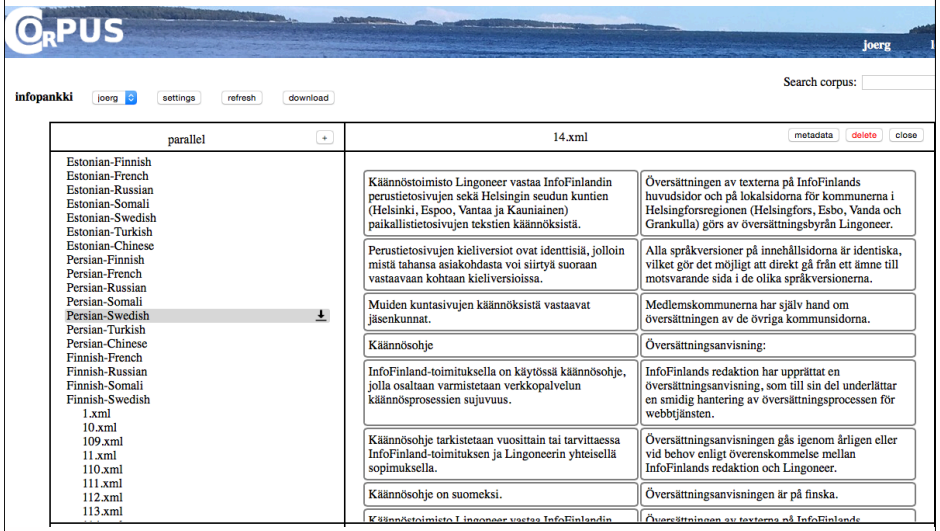
- data collection
- data extraction / conversion
- cross-lingual alignment
- data management

OPUS
corpus repository

MT data management system



MT data management system



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

TILDE

Tilde Custom Machine Translation

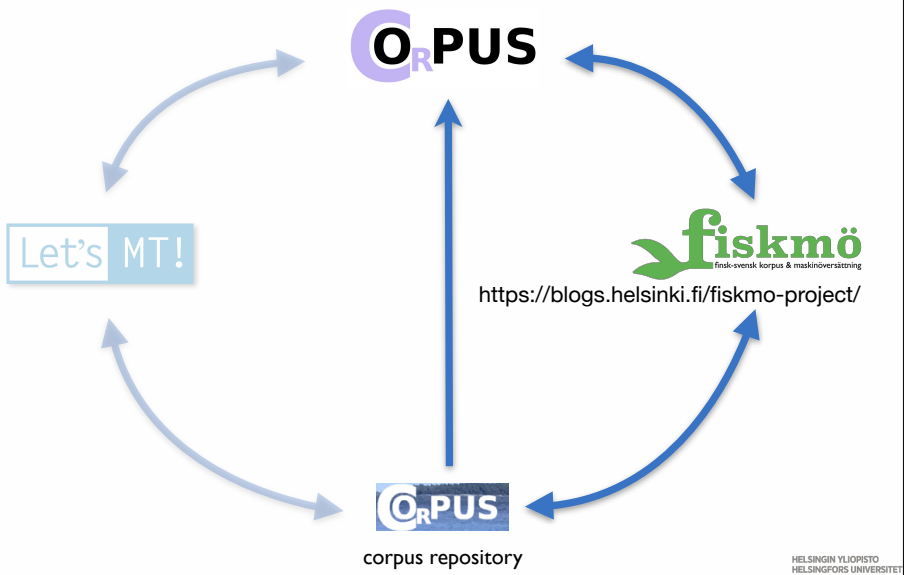
Learn more →

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

(3) The fiskmö project



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

Svenska kulturfonden [kites] HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI TURUN YLIOPISTO

fiskmö
finsk-svensk korpus & maskinöversättning

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

- alignment
- data curation
- machine translation
- user interfaces

Jörg Tiedemann

Mikko Aulamo

Svenska kulturfonden [kites] HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI TURUN YLIOPISTO

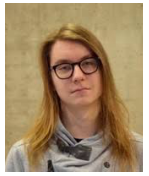
fiskmö
finsk-svensk korpus & maskinöversättning

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

Tommi Nieminen

Niko Papula

Matias Tamminen



- user data collection
- workflow integration

Svenska kulturfonden [kites] HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI TURUN YLIOPISTO

fiskmö
finsk-svensk korpus & maskinöversättning

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

- web crawling
- bitext extraction

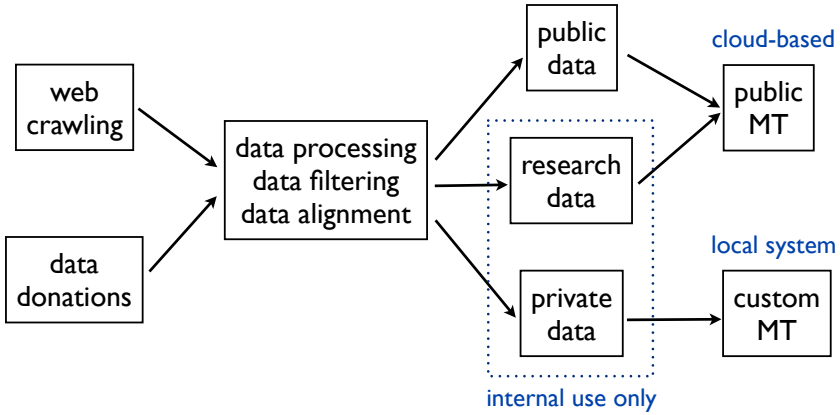
Filip Ginter

Jenna Kanerva

Akseli Leino

Svenska kulturfonden [kites] HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI TURUN YLIOPISTO

How does it work?



Svenska kulturfonden

What are the benefits?

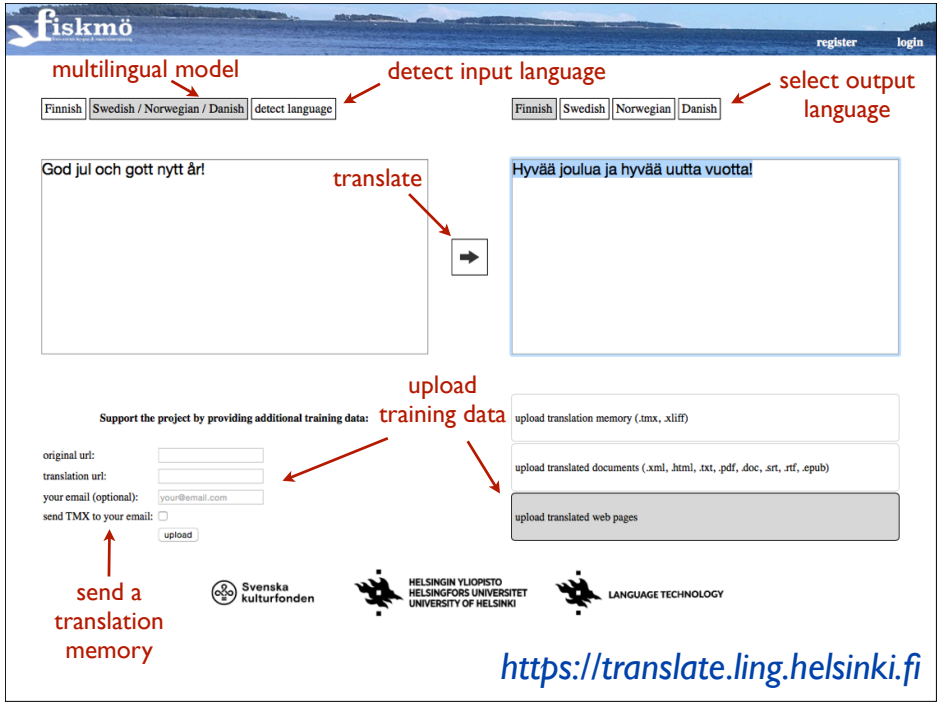
Why should you support us and what is in there for you?

- 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!

Svenska kulturfonden

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



https://translate.ling.helsinki.fi

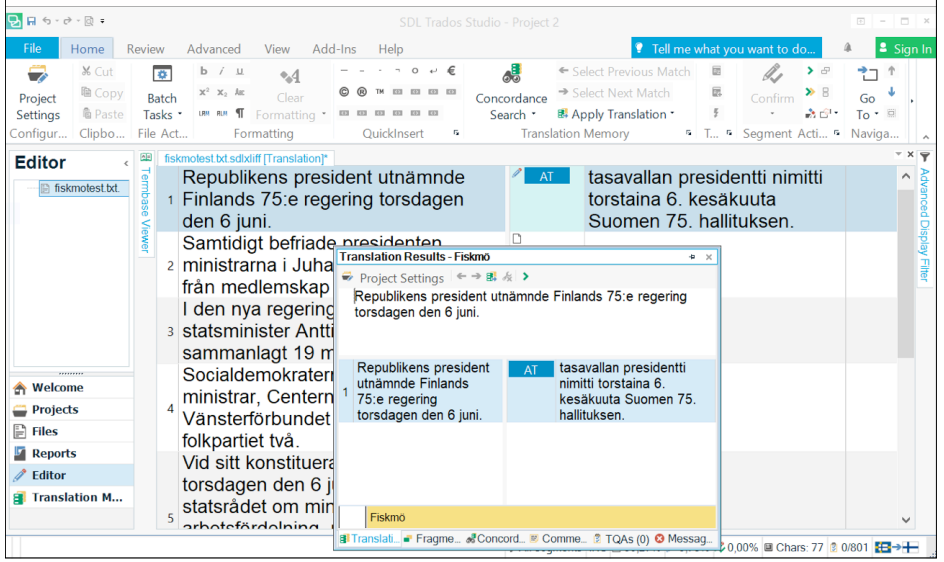
Available data (translation memories)

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

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

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Embedded fiskmö MT in Trados Studio



How good is it now?

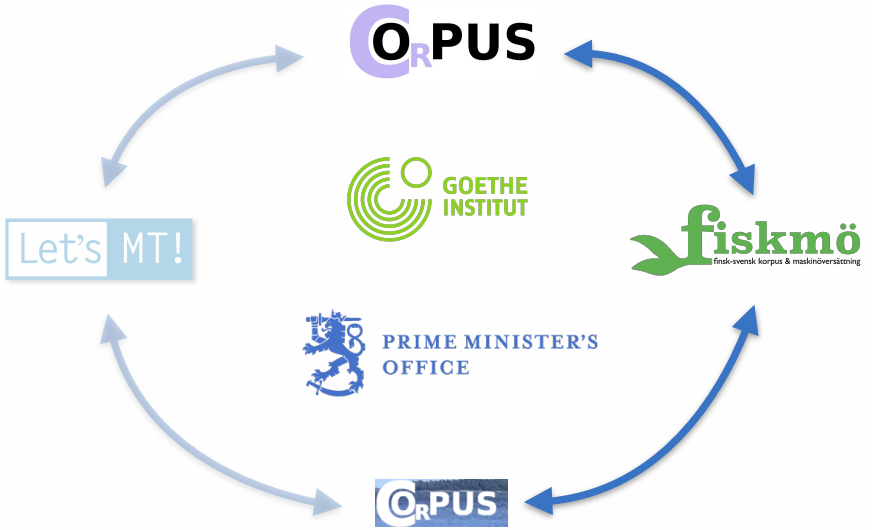
A recent result with our fiskmö benchmark test set:

	BLEU	chr-F
• from Finnish to Swedish		
• Google Translate	17.7	0.513
• Presidency MT	**29.5**	**0.627**
• 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	**26.1**	**0.623**

Svenska kulturfonden

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

(4) Other Collaborations



corpus repository

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

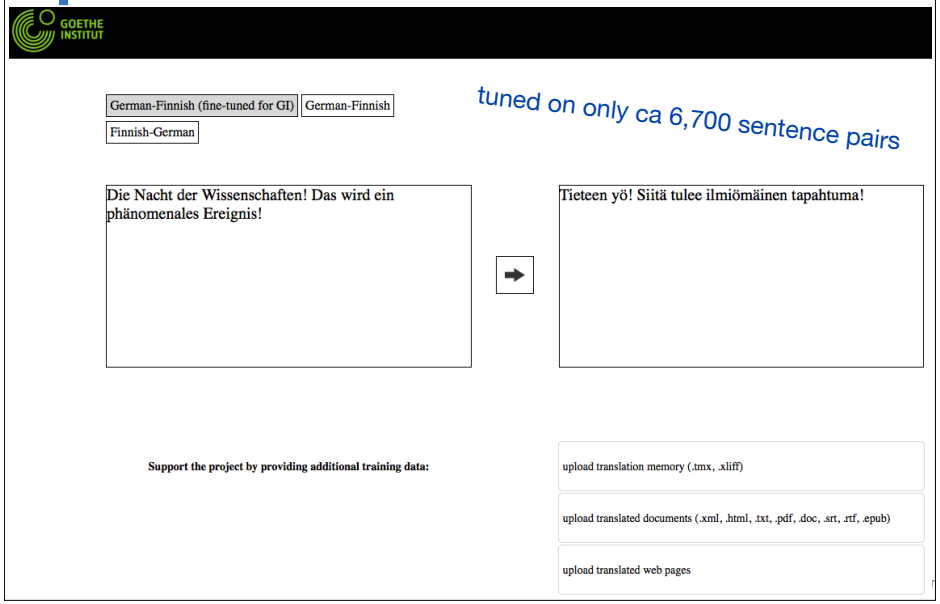
Data collection

- 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



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*
Helsinki - tuned	***21.4***

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

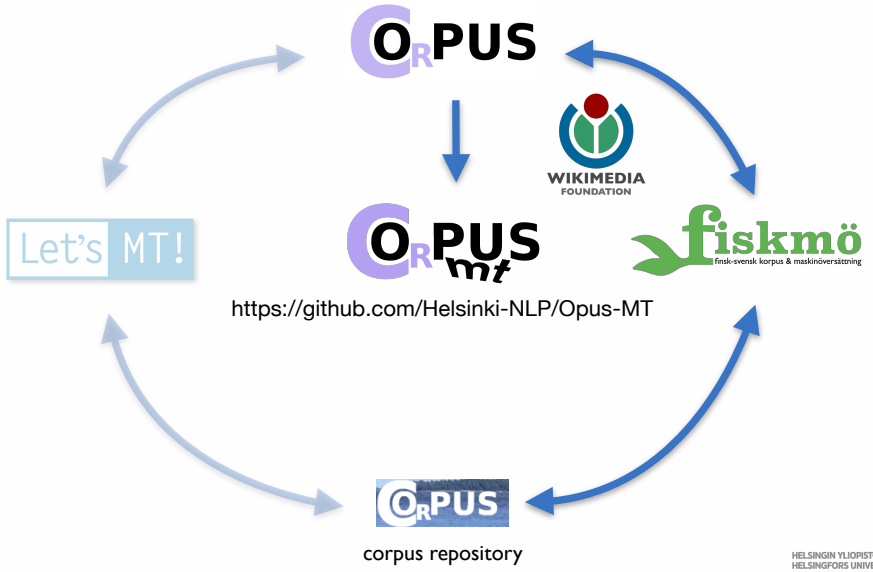
Outlook into the future

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



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

corpus repository

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

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)

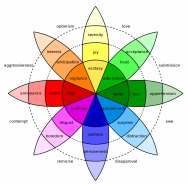
- de-centralized user-contributed translation services
- automatic load balancing
- personalized services (with feedback loops)

Related research in Helsinki

Data collection for MT
> 200 languages

OPUS

http://opus.nlpl.eu



cross-lingual sentiment analysis

FOTRAN
Found in Translation

learning semantics
from > 1,000 languages

audiovisual data & MT
6 languages

MeMAD
Methods for Managing Audiovisual Data

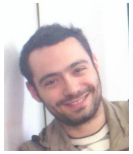
https://memad.eu

fiskmö
finsk-svensk korpus & maskinöversättning

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

Language Technology in Helsinki

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

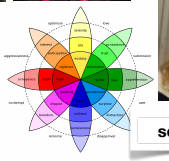
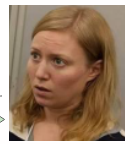

NLP for Finno-Ugric languages


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*

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI