European Language Grid

An Introduction and Overview
Katrin Marheinecke (DFKI) – ELG Project Manager (katrin.marheinecke@dfki.de)
15-12-2020 – 4th Regional ELG Workshop – Finland
http://www.european-language-grid.eu
Coordinator: Dr. Georg Rehm (DFKI)
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00</td>
<td>Welcome and introduction</td>
<td>Krister Lindén (University of Helsinki)</td>
</tr>
<tr>
<td>14:05</td>
<td>ELG: history and overview</td>
<td>Katrin Marheinecke (DFKI)</td>
</tr>
<tr>
<td>14:30</td>
<td>ELG: online demo</td>
<td>Nils Feldhus (DFKI)</td>
</tr>
<tr>
<td>14:50</td>
<td>Finnish Pilot Projects funded in ELG</td>
<td>Filip Ginter (University of Turku)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sebastian Andersson (Lingsoft)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jörg Tiedemann (University of Helsinki)</td>
</tr>
<tr>
<td>15:20</td>
<td>Finnish Language Platform Initiative</td>
<td>Marko Turpeinen (1001 Lakes)</td>
</tr>
<tr>
<td>15:40</td>
<td>Summary and discussion</td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>End of the workshop</td>
<td></td>
</tr>
</tbody>
</table>
Point of Departure: Multilingualism in Europe

- Multilingualism is at the heart of the European idea
- 24 official EU languages – they all have the same status
- Dozens of co-official, regional and minority languages as well as languages of immigrants and trade partners
- Many economic, social and technical challenges
  - The Digital Single Market needs to be multilingual
  - Cross-border, cross-lingual, cross-cultural communication
  - Fragmentation of the LT market and landscape
Language Technology is already everywhere

- Language Technology makes use of theoretical results of language-oriented research to create applications and technological solutions. Fields and areas involved:
  - Artificial Intelligence + Computer Science
  - Computational Linguistics
    - Natural Language Processing (NLP)
    - Natural Language Understanding (NLU)
  - Psychology, Psycholinguistics
  - Cognitive Science
- **Language Technology = Language-centric AI**

  - Spell and grammar checker in MS Word
  - Web search (Google, Bing, Yandex etc.)
  - Social Media Analytics, Media Monitoring
  - Voice control for phones or computers
  - Voice control in cars
  - Machine translation
  - Conversational agents and chatbots (Echo, Home, Siri, Cortana etc.)
  - Speech synthesis in games
  - Computer-Assisted Language Learning (CALL)
  - Optical Character Recognition (OCR)

Selected applications that include Language Technology
Motivation and Context

• The European LT landscape and market are very fragmented.

• The European LT community has been demanding a European Language Technology platform for several years.

• META=NET vision: European Service Platform for Language Technologies.

• First mention in the Strategic Research Agenda for Multilingual Europe 2020 (published in early 2013).
European LT Market – CEF Study

Approach and Observations

• Background: SMART 2016/0103 contract: contribute to CEF AT as “multilingualism enabler” for CEF DSIs.

• European LT vendors grouped per type of tech: Translation, Speech, Search, NLU, Analytics

• EU market approx. 1B€ in 2020 – disrupted by dominant global players

• SMEs: 70% of EU LT vendors up to 50 employees

• Revenue per company is growing

• Market is highly fragmented: hundreds of SMEs, many address very specific niches, sectors and languages

Recommendations

• Europe is strong in R&I, but not successful to scale innovations and capture the market

• Europe needs European alternatives to fill the gaps and deficiencies and to avoid reliance on monopolies

• Multilingual DSM should be developed on its own infrastructure

• Public procurement can be the major driver for European LT industry to avoid dependence on monopolies

• Plans needed to avoid brain drain

➤ A platform is needed to connect demand and supply as well as industry and research

➤ CEF study estimates the European LT Market to reach 1B€ in 2020.

The Global Natural Language Processing Market size is expected to react $29.5 billion by 2025 rising at a market growth of 20.5% CAGR during the forecast period

(Sourcr: Global Newswire)
European Language Grid

ELG Project
Objectives (Selection)

1. Establish the ELG as the primary Language Technology platform and market place in Europe to tackle the fragmentation of the European LT landscape.

2. ELG as a platform for commercial and non-commercial, industry-related LTs, both functional and non-functional.

3. Enable the European LT community to upload services and data sets into the ELG, to deploy them and to connect with, and make use of those resources made available by others.

4. Enable businesses to grow and benefit from scaling up.

5. Unleash enormous potential for innovation.
European Language Grid – Current State of Play (December 2020)

- User registration, authentication, authorisation
- User categories, respective rights and policies
- LT metadata upload and editing facilities
- Metadata conversion and harvesting
- LT service registration, integration
- LT service try out and execution
- LT data browsing, searching, upload, download
- Online documentation

- Current state of play in the European Language Grid towards ELG Release 2 (early 2021):
  - 168 functional services and tools
  - 1998 corpora
  - 729 lexical/conceptual resources
  - 7 language descriptions

- Facilities for consumers, esp. of data and services
- Facilities for providers, esp. of data and services

Users can connect to the ELG cloud platform via ELG APIs, remote APIs, ELG GUI, Python client library, download of containers or source code.
European Language Grid – What can you do with it?

**Data consumers** can search and browse the ELG catalogue
- for different types of data, language processing services, projects and organisations
- download data (depending on access conditions)

**Service consumers** can try out and test language processing services
- call a service from the command line, integrate it into workflow
- view code samples
- current API support: MT, IE, ASR, TTS, text classification

**Data and service consumers** can use a Python-based API for accessing the ELG catalogue, searching and directly fetching datasets to feed them into, e.g., their model training pipeline

**Data providers** can upload resources and register metadata descriptions using a semantic schema
- a dashboard provides an overview of own resources including their status (*draft* → *published*)

**Language Technology providers** can provide their service/tool as a Docker container
- Different options: Docker image directly in ELG; Docker image in ELG talks to remote service
LT Services: Current state of play

- **ELG Release 1** (April 2020) finalised APIs for major classes of services (ASR, IE, MT, TTS)
  - Concentrated on Czech, English, French, German, Greek, Latvian, Spanish (native ELG languages)
  - 150 distinct services for Text Analytics: 52 English, 28 German, 21 French, 14 Greek
  - 24 MT services
  - 9 ASR services
  - 2 TTS (Latvian, Lithuanian)

- **ELG Release 2** (early 2021) will add support for other EU and related languages, at least:
  - 8 further ASR, 200-250 further Text Analysis, 23 further MT, 9 further TTS services
  - We also expect the first services and datasets from the ELG pilot projects

- **ELG Release 3** (early 2022)
  - Services for an even wider range of non-EU languages, current projection is for at least
    - 15 further ASR, 160 further Text Analysis, 9 further MT services
  - Additional service types: image OCR, terminology extraction from corpora, etc.

<table>
<thead>
<tr>
<th>Language Group</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic linguistic processing</td>
<td>146</td>
<td>30</td>
<td>63</td>
<td>239</td>
</tr>
<tr>
<td>Entity recognition and linking</td>
<td>102</td>
<td>7</td>
<td>11</td>
<td>120</td>
</tr>
<tr>
<td>Classification (sentiment, opinion, topic, language etc.)</td>
<td>40</td>
<td>4</td>
<td>18</td>
<td>62</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>105</td>
<td>4</td>
<td>14</td>
<td>123</td>
</tr>
<tr>
<td>Other text processing (including parsing, summarisation etc.)</td>
<td>32</td>
<td>7</td>
<td>13</td>
<td>52</td>
</tr>
<tr>
<td>Speech tools (ASR, gender detection, synthesis etc.)</td>
<td>26</td>
<td>5</td>
<td>14</td>
<td>45</td>
</tr>
</tbody>
</table>

Services and tools to be ingested by the ELG partners (according to the proposal)
## ELG: Data Sets and Language Resources

<table>
<thead>
<tr>
<th>Language</th>
<th>Corpora</th>
<th>Lexical and Conceptual Resources</th>
<th>Models and Computational Grammars</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELRA</td>
<td>635</td>
<td>545</td>
<td>-</td>
<td>1180</td>
</tr>
<tr>
<td>ELRC-SHARE</td>
<td>844</td>
<td>43</td>
<td>-</td>
<td>887</td>
</tr>
<tr>
<td>META-SHARE</td>
<td>52</td>
<td>12</td>
<td>7</td>
<td>71</td>
</tr>
<tr>
<td>LINDAT/CLARIAH-CZ</td>
<td>243</td>
<td>66</td>
<td>-</td>
<td>309</td>
</tr>
<tr>
<td>ELRA-SHARE</td>
<td>46</td>
<td>25</td>
<td>-</td>
<td>71</td>
</tr>
<tr>
<td>Zenodo</td>
<td>36</td>
<td>37</td>
<td>-</td>
<td>73</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1857</td>
<td>727</td>
<td>7</td>
<td>2591</td>
</tr>
</tbody>
</table>

### Language Group A

<table>
<thead>
<tr>
<th></th>
<th>Corpora</th>
<th>Lexicons</th>
<th>Models</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>META-SHARE yes</td>
<td>617</td>
<td>447</td>
<td>16</td>
<td>1323</td>
</tr>
<tr>
<td>META-SHARE no</td>
<td>582</td>
<td>550</td>
<td>1</td>
<td>1534</td>
</tr>
<tr>
<td>ELRA-SHARE yes</td>
<td>317</td>
<td>114</td>
<td>0</td>
<td>431</td>
</tr>
<tr>
<td>ELRA-SHARE no</td>
<td>74</td>
<td>16</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>ELDA no</td>
<td>563</td>
<td>1012</td>
<td>0</td>
<td>1932</td>
</tr>
<tr>
<td>ELG mixed</td>
<td>74</td>
<td>108</td>
<td>45</td>
<td>263</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2227</td>
<td>2247</td>
<td>60</td>
<td>5582</td>
</tr>
</tbody>
</table>

### Language Group B

<table>
<thead>
<tr>
<th></th>
<th>Corpora</th>
<th>Lexicons</th>
<th>Models</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>META-SHARE yes</td>
<td>617</td>
<td>447</td>
<td>16</td>
<td>1323</td>
</tr>
<tr>
<td>META-SHARE no</td>
<td>582</td>
<td>550</td>
<td>1</td>
<td>1534</td>
</tr>
<tr>
<td>ELRA-SHARE yes</td>
<td>317</td>
<td>114</td>
<td>0</td>
<td>431</td>
</tr>
<tr>
<td>ELRA-SHARE no</td>
<td>74</td>
<td>16</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>ELDA no</td>
<td>563</td>
<td>1012</td>
<td>0</td>
<td>1932</td>
</tr>
<tr>
<td>ELG mixed</td>
<td>74</td>
<td>108</td>
<td>45</td>
<td>263</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2227</td>
<td>2247</td>
<td>60</td>
<td>5582</td>
</tr>
</tbody>
</table>

### Language Group C

<table>
<thead>
<tr>
<th></th>
<th>Corpora</th>
<th>Lexicons</th>
<th>Models</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>META-SHARE yes</td>
<td>617</td>
<td>447</td>
<td>16</td>
<td>1323</td>
</tr>
<tr>
<td>META-SHARE no</td>
<td>582</td>
<td>550</td>
<td>1</td>
<td>1534</td>
</tr>
<tr>
<td>ELRA-SHARE yes</td>
<td>317</td>
<td>114</td>
<td>0</td>
<td>431</td>
</tr>
<tr>
<td>ELRA-SHARE no</td>
<td>74</td>
<td>16</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>ELDA no</td>
<td>563</td>
<td>1012</td>
<td>0</td>
<td>1932</td>
</tr>
<tr>
<td>ELG mixed</td>
<td>74</td>
<td>108</td>
<td>45</td>
<td>263</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2227</td>
<td>2247</td>
<td>60</td>
<td>5582</td>
</tr>
</tbody>
</table>

### Language Resources

- **Official EU Languages**: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish
- **EU candidate languages and Free Trade Partners**: Albanian, Basque, Catalan, Galician, Icelandic, Norwegian, Scottish Gaelic, Welsh, Serbian, Turkish, Ukrainian
- **Languages spoken by EU immigrants; languages of political and trade partners**: Afrikaans, Arabic, Berber, Cebuano, Chinese, Hebrew, Hindi/Urdu, Indonesian, Japanese, Korean, Kurdish, Latin, Malay, Pashto, Persian (Farsi), Russian, Tamil, Vietnamese

Datasets and resources to be provided by the ELG partners (according to the proposal)
Stakeholders and Users

Companies that
- ... develop Language Technologies
- ... integrate Language Technologies
- ... purchase Language Technologies

Universities and research centres that
- ... develop Language Technologies
- ... use Language Technologies

Public administrations that purchase or use Language Technologies

Other organisations (e.g., NGOs) that purchase or use Language Technologies

Funding agencies that support the development of Language Technologies
META-FORUM Conference Series

META-FORUM 2020 – December 01-03, virtual conference
  Piloting the European Language Grid
META-FORUM 2019 – October 08/09, Brussels, Belgium
  Introducing the European Language Grid
META-FORUM 2017 – November 13/14, Brussels, Belgium
  Towards a Human Language Project
META-FORUM 2016 – July 04/05, Lisbon, Portugal
  Beyond Multilingual Europe
META-FORUM 2015 – April 27, Riga, Latvia
  Technologies for the Multilingual Digital Single Market
META-FORUM 2013 – September 19/20, Berlin, Germany
  Connecting Europe for New Horizons
META-FORUM 2012 – June 20/21, Brussels, Belgium
  A Strategy for Multilingual Europe
META-FORUM 2011 – June 27/28, Budapest, Hungary
  Solutions for Multilingual Europe
META-FORUM 2010 – November 17/18, Brussels, Belgium
  Challenges for Multilingual Europe
Community: NCCs and LTC

32 National Competence Centres (NCCs)

- Strong international network of national networks to broaden ELG’s reach, identify content for the ELG and interest companies in using the ELG.
- Main goal: *support the mission of the ELG project.*

**European LT Council (LTC)**

- A pan-European body, in which strategic LT-related matters can be discussed and coordinated.
- Main goal: *represent and support European LT community.*
1. **Austria**: Dagmar Gromann, Zentrum für Translationswissenschaft, Universität Wien
2. **Belgium**: Walter Daelemans, Comp. Ling. and Psycholing. Res. Centre (CLIPS), University of Antwerp
3. **Bulgaria**: Svetla Koeva, Bulgarian Academy of Sciences
4. **Croatia**: Marko Tadic, Inst. of Linguistics, Faculty of Hum. and Social Science, University of Zagreb
5. **Cyprus**: Dora Loizidou, University of Cyprus
6. **Czech Republic**: Jan Hajic, Inst. of Formal and Applied Linguistics, Charles University in Prague
7. **Denmark**: Bolette Sandford Pedersen, Centre for Lang. Technology, Department of Nordic Research, University of Copenhagen
8. **Estonia**: Kadri Vare, Department of Language, Estonian Ministry of Education and Resear
9. **Finland**: Krister Lindén, Department of Modern Languages, University of Helsinki
10. **France**: François Yvon, CNRS-LIMSI
11. **Germany**: Georg Rehm, Speech and Language Technology Lab, DFKI
12. **Greece**: Maria Gavrilliou, ILSP, R.C. “Athena”
13. **Hungary**: Tamás Várárdi, Research Institute for Linguistics, Hungarian Academy of Sciences
14. **Iceland**: Eirikur Rögnvaldsson, School of Humanities, University of Iceland
15. **Ireland**: Andy Way, ADAPT Centre and School of Computing, Dublin City University
16. **Italy**: Bernardo Magnini, Human Language Technology, Fondazione Bruno Kessler (FBK)
17. **Latvia**: Inguna Skadina, Institute of Mathematics and Computer Science, University of Latvia
18. **Lithuania**: Albina Auksošiūtė, Institute of the Lithuanian Language
19. **Luxembourg**: Dimitra Anastasiou, Luxembourg Institute of Science and Technology
20. **Malta**: Michael Rosner, Department Intelligent Computer Systems, University of Malta
21. **Netherlands**: Jan Odijk, Utrecht Institute of Linguistics, Universiteit Utrecht
22. **Norway**: Kristine Eide, The Language Council of Norway – Språkrådet
23. **Poland**: Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
24. **Portugal**: António Branco, Department of Informatics, University of Lisbon
25. **Romania**: Dan Tufis, Research Institute for Artificial Intelligence, Romanian Academy of Sciences
26. **Serbia**: Cvetana Krstev, Faculty of Mathematics, Belgrade University (UBG)
27. **Slovakia**: Radovan Garabík, Ludovit Stur Institute of Linguistics, Slovak Academy of Sciences
28. **Slovenia**: Simon Krek, Jožef Stefan Institute
29. **Spain**: Marta Villegas, Barcelona Supercomputing
30. **Sweden**: Jens Edlund, Speech, Music & Hearing/Språkbanken Tal, KTH Royal Institute of Technology
31. **Switzerland**: Hervé Bourlard, Idiap Research Institute
32. **UK**: Kalina Bontcheva, Department of Computer Science, University of Sheffield
European Language Grid: Sustainable Operational Model & Legal Entity

- ELG is supposed to be a long-term, sustainable initiative – a legal entity is needed.
- The technical and operational requirements – high availability and performance, SLAs, billing, support etc. – create non-trivial costs: hosting; bandwidth; ELG team; legal; etc.
- We’ve identified several ways and approaches of covering the costs on a long-term basis.
- Establish consensus for a sustainable operational model.
- Options: a) for-profit or b) not-for-profit company, c) association, d) foundation
- First concept prepared (Business Model Canvas).
- Q4/2021: Establish legal entity (probably with a soft start)
Open Calls for Pilot Projects

Two open calls for pilot projects
- **Open Call #1**: 03/04 2020
- **Open Call #2**: 10/11 2020

Pilot projects shall
- **Type A**: broaden ELG’s portfolio or
- **Type B**: demonstrate usefulness of ELG

Up to €200,000 per project

Approx. €2,000,000 FSTP in total

Available in Open Call #1: 1.3M€
Available in Open Call #2: 585k€

Project duration: 9-12 months
Eligibility: SMEs, research orgs.

Open Call #1 – Statistics

110 project proposals evaluated
- **Type A**: 79 proposals
- **Type B**: 31 proposals

Total amount requested: 16.9M€
10 projects selected on 29 June 2020

Open Call #2 – Statistics

106 project proposals submitted
## Open Call #1: Selected Pilot Projects

<table>
<thead>
<tr>
<th>Organization</th>
<th>Project Description</th>
<th>Country</th>
<th>Funding (EUR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fondazione Bruno Kessler</td>
<td>European Clinical Case Corpus</td>
<td>Italy</td>
<td>139,370</td>
</tr>
<tr>
<td>Lingsoft, Inc.</td>
<td>Lingsoft Solutions as Distributable Containers</td>
<td>Finland</td>
<td>140,625</td>
</tr>
<tr>
<td>Coreon GmbH</td>
<td>MKS as Linguistic Linked Open Data</td>
<td>Germany</td>
<td>167,375</td>
</tr>
<tr>
<td>Elhuyar Fundazioa</td>
<td>Basque-speaking smart speaker based on Mycroft AI</td>
<td>Spain</td>
<td>117,117</td>
</tr>
<tr>
<td>Universita’ Degli Studi di Torino</td>
<td>Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools [...]</td>
<td>Italy</td>
<td>126,125</td>
</tr>
<tr>
<td>University of Helsinki</td>
<td>Open Translation Models, Tools and Services</td>
<td>Finland</td>
<td>154,636</td>
</tr>
<tr>
<td>Centre for Translation Studies,</td>
<td>Extracting Terminological Concept Systems from Natural Language Text</td>
<td>Austria</td>
<td>132,977</td>
</tr>
<tr>
<td>University of Vienna</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Turku, Turku NLP</td>
<td>Textual paraphrase dataset for deep language modelling</td>
<td>Finland</td>
<td>166,085</td>
</tr>
<tr>
<td>research group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weber Consulting KG</td>
<td>Virtual Personal Assistant Prototype</td>
<td>Austria</td>
<td>87,445</td>
</tr>
<tr>
<td>FZI Research Center for Information</td>
<td>Streaming Language Processing in Manufacturing</td>
<td>Germany</td>
<td>132,160</td>
</tr>
<tr>
<td>Technology</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
European Language Grid

https://www.european-language-grid.eu

• ELG project website
• Typical EU project information plus a lot of additional content (events, open calls, community, NCCs etc.)
• The whole content of the project website will be integrated into the European Language Grid website (proper) in 2021

https://live.european-language-grid.eu

• The actual European Language Grid platform and website
• The next revision of the ELG project website (left) will include dynamic content that will be pulled from the ELG repository (right), e.g., languages, number of services, categories of services etc.
Future roles of the EUROPEAN LANGUAGE GRID

Exploitation and dissemination channel for future Horizon Europe, DEP and other projects

Main platform and marketplace of the European LT community

Bridge between the Language-centric AI and broader European AI community

Processing service and data portal for the Digital Europe Programme

Infrastructure for future Horizon Europe projects

Humane AI
Human-Centered Artificial Intelligence
Summary and Next Steps

- Establish ELG as the primary platform and marketplace for Language Technology in Europe.
- An initiative from the European LT community for the European LT community.
- European LT landscape is highly fragmented: ELG aims to provide just the right umbrella platform.
- Global market size by 2025 is enormous: we want the European LT community to be a key player.
- We want to increase the visibility and reach of all members of the European LT landscape.
- ELG is a long-term initiative: we will establish a legal entity for sustainability.
- Contribute to the long-term goal of Digital Language Equality in Europe by giving all our languages one virtual home and umbrella platform that collects all services and resources (ELE).
- **Next steps:** develop and populate ELG Release 2 with functional services and resources (early 2021); develop ELG Release 3 (early 2022); establish the ELG legal entity (late 2021).
EUROPEAN LANGUAGE GRID

Interested to learn more?

• Visit our project website:
  https://www.european-language-grid.eu

• Go to the live ELG:
  https://live.european-language-grid.eu

• Like to have a user account for the grid? Send an email to:
  contact@european-language-grid.eu
Thank you!

Interested in participating?
Please get in touch:
contact@european-language-grid.eu

Katrin Marheinecke (DFKI, ELG Project Manager)
katrin.marheinecke@dfki.de

15-12-2020 – 4th Regional ELG Workshop – Finland
http://www.european-language-grid.eu

Coordinator: Dr. Georg Rehm (DFKI)
OPUS-MT - Public NMT Models & Tools
Language Technology - University of Helsinki

Available software:
• MT server solution based on Marian NMT
• dockerised web-app, translation interface and API
• NMT training pipeline (OPUS-MT-train)
• CAT integration (OPUS-CAT)

Pre-trained translation models:
• number of bilingual models: 1,048
• number of multilingual models: 53
• number of supported source languages: 229
• number of supported target languages: 222
• number of supported language pairs: 1,715

Protoytype for Sámi language NMT
https://translate.ling.helsinki.fi/ui/sami

Embedded OPUS-MT in CAT tools
https://github.com/Helsinki-NLP/OPUS-CAT
Embedded OPUS-MT in CAT tools

https://github.com/Helsinki-NLP/OPUS-CAT

Tornado web-app server

Configuration

The server.py program accepts a configuration file in json format. By default it try to use config.json in the current directory. But you can give a custom one using --conf flag.

An example configuration file looks like this:

```json
{
  "en": {
    "es": {
      "configuration": "/models/en-es/decoder.yaml",
      "host": "localhost",
      "port": "18082"
    },
    "fr": {
      "configuration": "/models/en-fr/decoder.yaml",
      "host": "localhost",
      "port": "18082"
    }
  }
}
```

This example configuration can provide MT service for en→es and en→fr language pairs.

- configuration: points to a yaml file containing the decoder configuration usable by marian-server. If this value is not provided, Opus-MT will assume that the service is already running in a remote host and port as given in other options. If value is provided a new sub process will be created using marian-server.
- host: The host where the server is running.
- port: The port to be listen for marian-server.

Websocket client: Token-aligned output

```
echo "Mitä kului? Mäkentäs on hauskaa." ./opusMT-client.py -H localhost -p 2808 -s fi -t en
```

This should return something like

```json
{
  "alignment": [
    "0-0 0-2 1-1 2-3",
    "0-0 1-3 2-4 3-5 4-6"
  ],
  "result": "How are you? The translation is fun.",
  "source": "fi",
  "source-sentences": [
    "Mitä kului?", 
    "Mäkentäs on hauskaa."
  ],
  "target": "en",
  "target-sentences": [
    "How are you?",
    "The translation is fun."
  ]
}
```
Train Opus-MT models

This package includes scripts for training NMT models using MerianMT and OPUS data for OPUS-MT. More details are given in the Makefile but documentation needs to be improved. Also, the targets require a specific environment and right now only work well on the CSC HPC cluster in Finland.

Pre-trained models

The subdirectory models contains information about pre-trained models that can be downloaded from this project. They are distributed with a CC-BY 4.0 license.

Quickstart

Setting up:

```
git clone https://github.com/Helsinki-NLP/OPUS-MT-train.git
git submodule update --init --recursive --remote
make install
```

Training a multilingual NMT model (Finnish and Estonian to Danish, Swedish and English):

```
make SRCLANGS="fi et" TRGLANGS="da sv en" train
make SRCLANGS="fi et" TRGLANGS="da sv en" eval
make SRCLANGS="fi et" TRGLANGS="da sv en" release
```

Development demo:

https://translate.ling.helsinki.fi/kaam
Provider of language services and language technology solutions

30+ years experience of NLP development for text and speech

Large collaboration network with leading universities in the Nordics
LSDISCO - Lingsoft Solutions as Distributable Containers

- We will provide access to our speech and language technology tools through ELG to organisations and companies with language needs for the Nordic languages.
- We will make our solutions easily distributable with scalable integration in other environments and solutions
- We will make them available both for commercial and non-commercial usage
Lingsoft Solutions

<table>
<thead>
<tr>
<th>Text Analysis</th>
<th>Speech Recognition</th>
<th>Machine Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemmatization, PoS tagging</td>
<td>Transcription</td>
<td>Customisation</td>
</tr>
<tr>
<td>NER and ontologies</td>
<td>Diarisation</td>
<td>Terminologies</td>
</tr>
<tr>
<td>Spelling and grammar</td>
<td>Subtitling</td>
<td>Professional CAT support</td>
</tr>
</tbody>
</table>

Languages supported:
- FI
- SV
- DA
- NB
- NN
- EN
- DE
Some Use Cases
Lingsoft on täydellinen palvelu kieltaloihin: Suomen suurin suomalainen käännöstämisto ja

Pohjoismaiden johtava kieli palvelujen ja kieliteknologiatautaisojen toimittaja.

Tarjoamme hetkeiset ja luotettavat käännöspalvelut, saavutettavat videotekstit

tehokkaat tekstinkäsittelypalvelut sekä monipuoliset puheentunnistus- ja tek.

Pastey

Ctrl+V
Machine Translation for Professional Translators

What is Lingsoft MT Lab?
## Metadata and Entity Recognition

### Lingsoft Analyser Demo

<table>
<thead>
<tr>
<th>Text</th>
<th>Analysis</th>
<th>CoNLL-U</th>
<th>JSON output</th>
<th>Unknowns</th>
<th>Terms</th>
<th>Ontology</th>
<th>Annotation</th>
</tr>
</thead>
</table>

Lingsoft on täyden palvelun kielitalo: Suomen suurin suomalainen käännöstoimisto ja Pohjoismaiden johtava kielipalvelujen ja kieliteknologiaratkaisujen toimittaja. Tarjoamalla Suomi (LOC) ja luotettavat käännöspalvelut, saavutettavat videotekstitykset, tehokkaat tekstinkäsittelypalvelut sekä monipuoliset pääemännistysten ja tekstianalytiikkaratkaisut.