

***Velkommen til  
Gå-hjem-møde  
om  
Statistisk baseret automatisk  
oversættelse***

18. April 2012

Center for Sprogteknologi  
Københavns Universitet



# Præsentation

## Center for Sprogteknologi, Københavns Universitet

- Forsker i automatisk oversættelse, semantiske ressourcer, multimodal kommunikation, maskinlæring
- Opbygger sprogteknologiske ressourcer og værktøjer
- Deltager i opbygning af en humanistisk forskningsinfrastruktur DIGHUMLAB, clarin.dk
- Udfører konsulentopgaver

## Oversættelsescentret, Københavns Universitet

- Oversættelse, tekstproduktion og sprogrevision inden for de fleste sprog



# Præsentation

Fra CST:

Chefkonsulent Hanne Fersøe [hannef@hum.ku.dk](mailto:hannef@hum.ku.dk)

Datalogvist Dorte Haltrup Hansen [dorteh@hum.ku.dk](mailto:dorteh@hum.ku.dk)

Videnskabelig medarbejder Sussi Olsen [saolsen@hum.ku.dk](mailto:saolsen@hum.ku.dk)

Seniorrådgiver Claus Povlsen [cpovlsen@hum.ku.dk](mailto:cpovlsen@hum.ku.dk)

Seniorforsker Jürgen Wedekind [jwedekind@hum.ku.dk](mailto:jwedekind@hum.ku.dk)

Seniorudviklingsingeniør Lene Offersgaard [leneo@hum.ku.dk](mailto:leneo@hum.ku.dk)

Fra Oversættelsescentret:

Annette Lindegaard og Kirsten Andersen, [oc@hum.ku.dk](mailto:oc@hum.ku.dk)



# Dagens program

15.00 – 15.10	Velkomst Seniorudviklingsingeniør Lene Offersgaard, Center for Sprogteknologi, Københavns Universitet
15.10 – 15.40	Præsentation af LetsMT! platformen og dens muligheder Seniorudviklingsingeniør Lene Offersgaard, Center for Sprogteknologi, Københavns Universitet
15.40 – 16.00	Brug af data på LetsMT-platformen Datalogvist Dorte H. Hansen og Videnskabelig medarbejder Sussi Olsen, Center for Sprogteknologi, Københavns Universitet
16.00 – 16.15	Kaffepause
16.15 – 16.45	Erfaringer med integration af SDL Trados og LetsMT! Annette Lindegaard og Kirsten Andersen, Oversættelsescentret, Københavns Universitet
16.45 –	Spørgsmål og debat Centret byder på en forfriskning



# Statistisk oversættelse

## Krav:

- Træning af oversættelsessystem, fx open source system
- Store datamængder til træning
- Kraftige computere
- Særlig opsætning
- Vedligehold af it og oversættelsessystem
- Udfordring at integrere med de værktøjer I allerede bruger

## En lettere tilgang:

Let's MT!





- ICT-PSP program: udvikling
- Øge tilgængelighed af automatisk oversættelse for *mindre sprog*
- Forsøge at lave et 'alternativ' til  **Google** oversæt
- *Skræddersyet* automatisk oversættelse
- Varighed 2½ år, afsluttes august 2012
-  vil involvere brugere



# Partnere:



[Tilde SIA](#), Letland



[University of Edinburgh, Human Communication Research Centre](#), UK



[University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics](#), Kroatien



[University of Copenhagen, Centre for Language Technology](#), Danmark



[Uppsala University, Department of Linguistics and Philology](#), Sverige



[Zoorobotics BV](#), Holland



[Moravia](#), Tjekkiet

# Visionen for LetsMT!

- Det skal være nemt at finde træningsdata
- Nemt at bruge sine egne dokumenter og TM'er til træning
- Nem adgang til computerkraft, uden at investere
- Ingen indviklet opsætning
- Integration med nogle værktøjer I allerede bruger

**Online Sharing of Training Data and Building User-Tailored Statistical Machine Translation Systems**



# letsmt.eu Platformen

## Oversætte

- To slags systemer: offentlige og private systemer
- Web-interface, SDL Trados, integreret i din browser

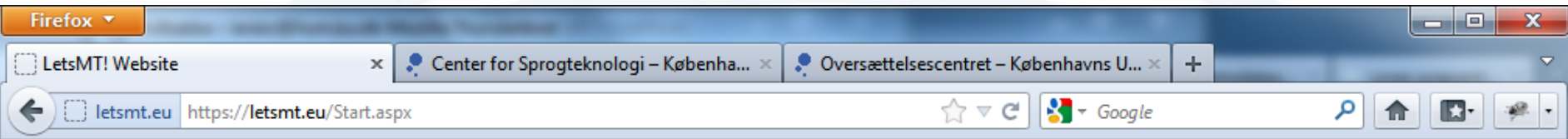
## Træning af systemer

- Bruge data der allerede er der
- Upload af egne data

## Repository med træningsdata

- Public og private data
- Public data kan bruges af alle til at træne systemer, og kun til træning! Private data er private
- NB: Træningsdata kan ikke downloades!



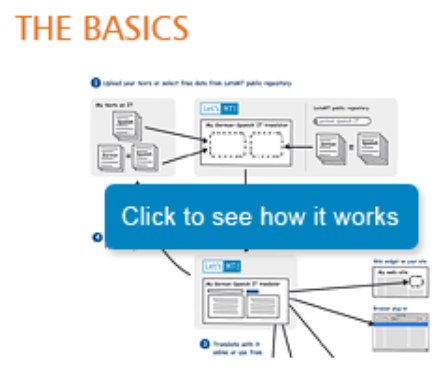


## Build your own machine translation system!

With LetsMT you can easily build and run your own custom machine translation systems. Simply upload your own corpora and/or choose to use any of the publicly available corpora. Train your systems and use them for all your translation needs.

- ### EASY
- ✓ Frees you from hardware and software infrastructure
  - ✓ Store corpora and engines in one place
  - ✓ Access to many free public texts and translation engines
  - ✓ Instantly increase productivity with CAT plug-in for localization process and web browser widget

- ### IS LETS MT FOR YOU?
- ✓ Localization & translation service providers
  - ✓ Holders of linguistic resources
  - ✓ Companies with a need to translate large amounts of information
  - ✓ Those who don't want to trust their resources to public systems



### Translate

Translate now using available systems

### Build

Build your own machine translation system

### Store and Share

Upload, organize, store, and share your corpora

# SMT systemer

Firefox

LetsMT! Website x Center for Sprogteknologi – Københa... x Oversættelsescentret – Københavns U... x +

letsmt.eu https://letsmt.eu/Systems.aspx

Let's MT! beta Systems Corpora Translate Tools About Administration

Welcome, ucph!  
Work in ucph

## Systems

Create system

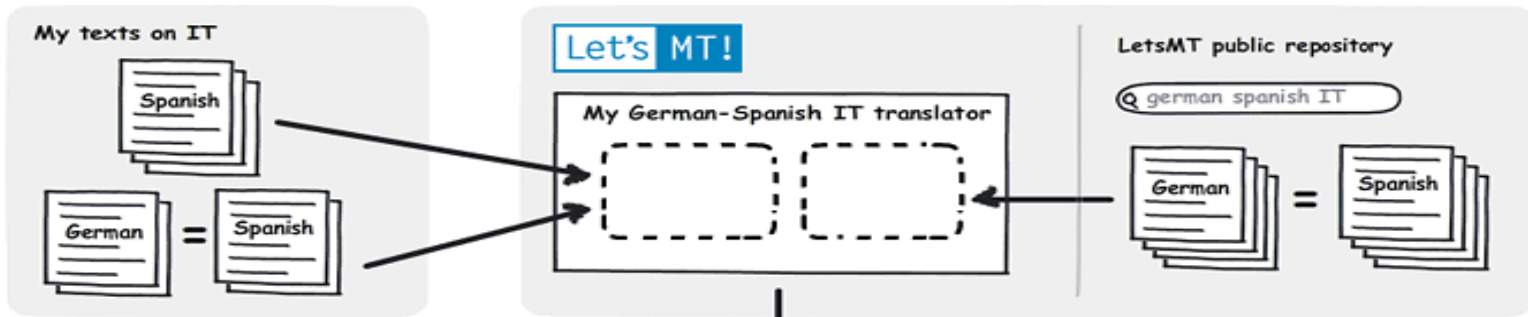
Source language: - select - Target language: Danish Owner (group): - select - Filter: - add filter -

Status	Name / Title	Source language	Target language	Subject Domain	Permissions
● Not Started	<a href="#">English - Danish Finance IV</a>	English	Danish	Other	Private
● Not Started	<a href="#">English-Danish Finance III</a>	English	Danish	Finance	Public
● Not Started	<a href="#">English-Danish Finance II</a>	English	Danish	Finance	Public
● Not Started	<a href="#">English-Danish Finance V</a>	English	Danish	Finance	Private
● Not Started	<a href="#">General English - Danish</a>	English	Danish	Other	Public

This is a list of public private statistical machine translation (SMT) systems.

An SMT system is an text translator you can train to translate general or specific subject domains.

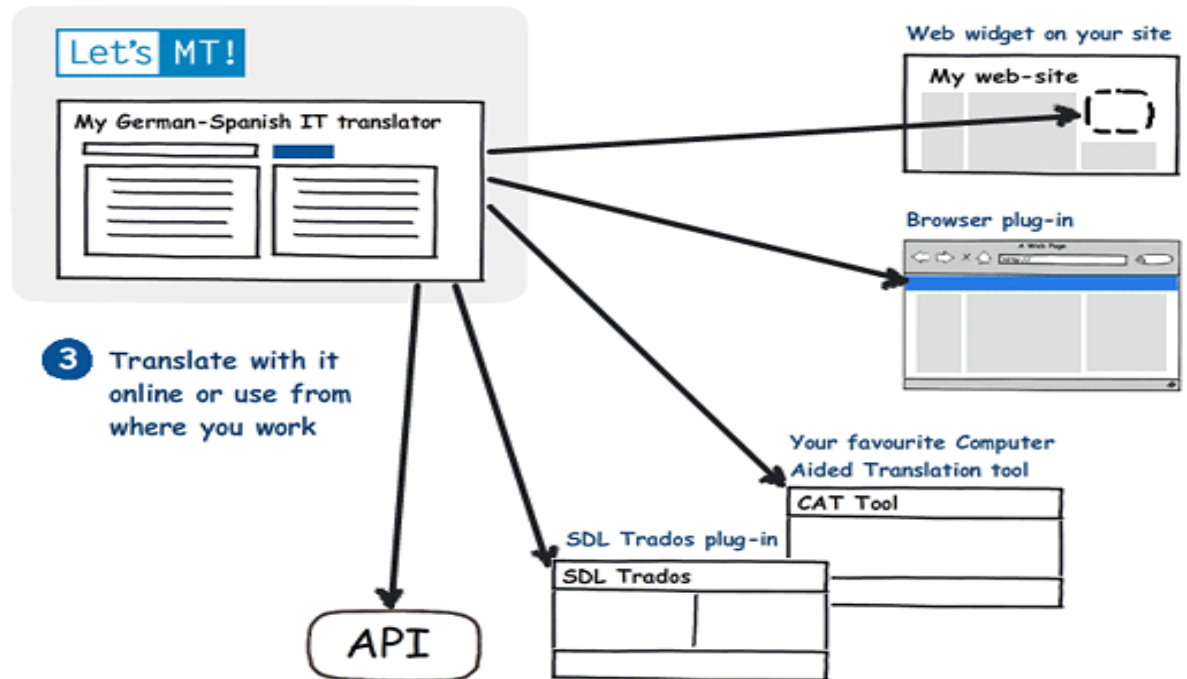
**1** Upload your texts or select free data from LetsMT public repository



**2** Create and train your system



**4** Add more texts for training



# SMT systemer

Firefox

LetsMT! Website × Center for Sprogteknologi – Københa... × Oversættelsescentret – Københavns U... ×

letsmt.eu https://letsmt.eu/Systems.aspx

Let's MT! beta Systems Corpora Translate Tools About Administration

Welcome, ucph!  
Work in ucph

## Systems

Create system

Source language: - select - Target language: Danish Owner (group): - select - Filter: - add filter -

Status	Name / Title	Source language	Target language	Subject Domain	Permissions
● Not Started	<a href="#">English - Danish Finance IV</a>	English	Danish	Other	Private
● Not Started	<a href="#">English-Danish Finance III</a>	English	Danish	Finance	Public
● Not Started	<a href="#">English-Danish Finance II</a>	English	Danish	Finance	Public
● Not Started	<a href="#">English-Danish Finance V</a>	English	Danish	Finance	Private
● Not Started	<a href="#">General English - Danish</a>	English	Danish	Other	Public

This is a list of public private statistical machine translation (SMT) systems.

An SMT system is an text translator you can train to translate general or specific subject domains.

# Detaljer om SMT systemer

The screenshot shows a Firefox browser window displaying the LetsMT! website. The address bar shows the URL <https://letsmt.eu/Systems.aspx>. The page title is "Systems". There are three tabs open: "LetsMT! Website", "Center for Sprogteknologi - Køben...", and "Oversættelsescentret - Københavns...".

At the top right, there is a blue button labeled "Create system". Below it, there are three dropdown menus for filtering: "Source language" (set to "- select -"), "Target language" (set to "Danish"), and "Owner (group)" (set to "- select -"). There is also a "- add filter -" dropdown.

Status	Name / Title	Source language	Target language	Subject Domain	Permissions
● Not Started	<a href="#">English - Danish Finance IV</a>	English	Danish	Other	Private
● Not Started	<a href="#">English-Danish Finance III</a>	English	Danish	Finance	Public

Below the table, there are two columns of details:

- Left Column:**
  - BLEU Score: 60,91
  - NIST Score: 9,5782
  - BLEU Score (CS): 59,83
  - NIST Score (CS): 9,4276
  - Monolingual corpus: 109 644 sentences
  - Parallel corpus: 113 897 sentences
  - Evaluation set: 1 000 sentences
  - Tuning set: 2 000 sentences
  - In-domain monolingual corpus: 109 644 sentences
- Right Column:**
  - Date Created: 2012.03.15 08:39:48
  - Date Modified: 2012.03.15 08:41:55
  - Date Accessed: 2012.03.15 08:41:55
  - Training started: 2012.03.15 08:42:01
  - Training finished: 2012.03.15 22:49:05
  - Owner (user): ucph
  - Parallel corpora: Semlab Business News 2, Semlab Business News 1
  - Monolingual corpora: Semlab Business News 2, Semlab Business News 1
  - In-domain parallel corpora: Semlab Business News 2, Semlab Business News 1
  - In-domain monolingual corpora: Semlab Business News 2, Semlab Business News 1

At the bottom, there are three buttons: "Details", "Start instance", and "View training chart".

## Machine translator

To use this online translation tool select one of the available SMT systems from the list.

System :

[Clear](#)

Translation finished

Enter the text you want to translate

Automatic translation results help to understand the meaning of the source text, but are not a substitute for a human translator

Name / Title: [University administration, KU 2b](#)

Source language: [Danish](#)

Target language: [English](#)

Subject Domain: [Education](#)

Permissions: [Private](#)

# Eksempler på oversættelser:

<http://maerskpress.com>

The Group's revenue increased by 7% to USD 60.2bn (USD 56.1bn), positively affected by higher oil prices and container volumes but offset by lower container freight rates.

Koncernens omsætning steg med 7% til USD 60.2bn (USD 56.1bn), positivt påvirket af højere oliepriser og containermængder men blev modsvaret af lavere containerfragtrater.

Profit was USD 3.4bn (USD 5.0bn) and thus in accordance with the latest announced outlook of USD 3.1-3.5bn expressed on 9 November 2011.

Resultatet var USD 3,4 mia. (USD 5,0 mia.) og på den måde, i overensstemmelse med den seneste udvikling på USD 3,1-3,5 mia. udtrykt den 9. november 2011.

The result was negatively affected by low container freight rates on especially the Asia-Europe trades and positively affected by the divestment gain from Netto Foodstores Limited, UK of USD 0.7bn.

Resultatet blev negativt påvirket af lav containerfragtrater på især Asien og Europa. positivt påvirket af salget af Netto Foodstores Limited i Storbritannien på USD 0,7 mia.

The Group's return on invested capital (ROIC) was 8.3% (12.2%).

Koncernens afkast af den investerede kapital (ROIC) var 8,3% (12,2%).

Available video includes b-rolls.

Disponible video omfatter b-rolls.





# Eksempler på oversættelser:

<http://www.nordea.com/Press/Press+releases>

The magazine Global Finance each year appoints the world's best investment banks in different categories. This year, in the category "Debt banks", Global Finance awarded Nordea as the best debt bank in the Nordic Region.

The editors of the magazine, with input from industry experts, used a series of criteria—including market share, number and size of deals, service and advice, structuring capabilities, distribution network, efforts to address market conditions, innovation, pricing, after-market performance of underwritings and market reputation—to select the winners.

Deals announced or completed in 2011 were considered.

Bladet Global Finance årligt udnævner verdens bedste investeringsbanker i forskellige kategorier. I år, i kategorien » kreditinstitutter Fordring awarded Nordea finansministers Global ", som den bedste gæld bank i Norden.

Den redaktører af input fra erhvervslivet, eksperter, der anvendes en række kriterier, deriblandt markedsandel, antal og størrelse af aftaler, service og rådgivning, strukturering kompetencer, distribution, indsats for at rette op på markedsvilkår, innovation, efter-marked, og resultaterne af underwritings — marked renommé udvælger vinderne.

Deals eller i 2011, blev anset for at være afsluttet.

# Træne et system

The screenshot shows a Firefox browser window displaying the LetsMT! website. The address bar shows the URL `https://letsmt.eu/System_step_1.aspx`. The page header includes the LetsMT! logo with a 'beta' tag, navigation links for 'Systems', 'Corpora', 'Translate', 'Tools', 'About', and 'Administration', and a user greeting 'Welcome, ucph!' with a 'Sign Out' link and a 'Work in' dropdown menu set to 'ucph'.

## SMT systems \ New system

If you have corpora (translation memories) that you would like to use you must [UPLOAD HERE](#) first and only then continue with system creation

### 1 System properties

Start creating your machine translation system by giving it a name and choosing source and target languages. The corpora (text data for training) which you will choose in the next steps will be filtered by the languages you have selected.

Name / Title \*

Source language \*

-select-

Target language \*

-select-

Permissions \*

-select-

-add field-

Save Save & Next Cancel

Check *Public* to make system available for translating to other LetsMT! users. They will not be able to modify anything, just use the system for translation.

# Træne et system II

Firefox
\_ [ ] X

LetsMT! Website
+

letsmt.eu
https://letsmt.eu/System\_step\_2\_1.aspx?id=smt-d553021e-4bf1-44ee-b262-adb0940f0b1e
☆ ↕ ↻
Google
🔍 🏠 🖨️

beta

Systems
Corpora
Translate
Tools
About
Administration

Welcome, ucph!
[Sign Out](#)

Work in
ucph

## SMT systems \ Lenex test

- Not Trained [Delete system](#)

1
System properties [Edit](#)

### Lenex test

🌐 English-Danish    🔒 Private

2
Parallel corpora

Select parallel corpora (texts) for training here.

🔍 Subject Domain
Finance
🔍 - add filter -

\* check D column if corpus is domain-specific and complies with current SMT system's domain. You can also include non domain-specific corpora to enrich the training set.

	D*	🔗 Name / Title	Subject Domain	Size	Permissions	Corpus Type
<input type="checkbox"/>	<input type="checkbox"/>	● Semlab Business News 1	Finance	0.2M	Public	Parallel
<input type="checkbox"/>	<input type="checkbox"/>	● Semlab Business News 2	Finance	0.2M	Public	Parallel
<input type="checkbox"/>	<input type="checkbox"/>	● Test eval	Finance	2k	Private	Parallel
<input type="checkbox"/>	<input type="checkbox"/>	● Eval_en-da_finance	Finance	1k	Public	Parallel

**Tips**

More parallel texts will be selected - better translation quality can be achieved.

---

**Recommended size of parallel corpora for training is at least 1 million sentences.**

# Træne et system III

The screenshot shows a Firefox browser window with the address bar displaying `https://letsmt.eu/System_step_default.aspx?id=smt-d553021e-4bf1-44ee-b262-adb0940f0b1`. The page title is "SMT systems \ Lenes test". Below the title, there is a radio button labeled "Not Trained". To the right of this section is a "Delete system" link. The main content area is a vertical list of six steps:

- 1 System properties** (with an "Edit" link)
- 2 Parallel corpora** (with a "Continue" link and a status "(Not completed)" with a yellow arrow icon)
- 3 Monolingual corpora**
- 4 Advanced options (optional)**
- 5 Training**
- 6 Evaluation**

Below the first step, the following details are visible:

- Lenes test**
- English-Danish
- Private

# Andre faciliteter

Upload af data, alignering af data, mere senere...

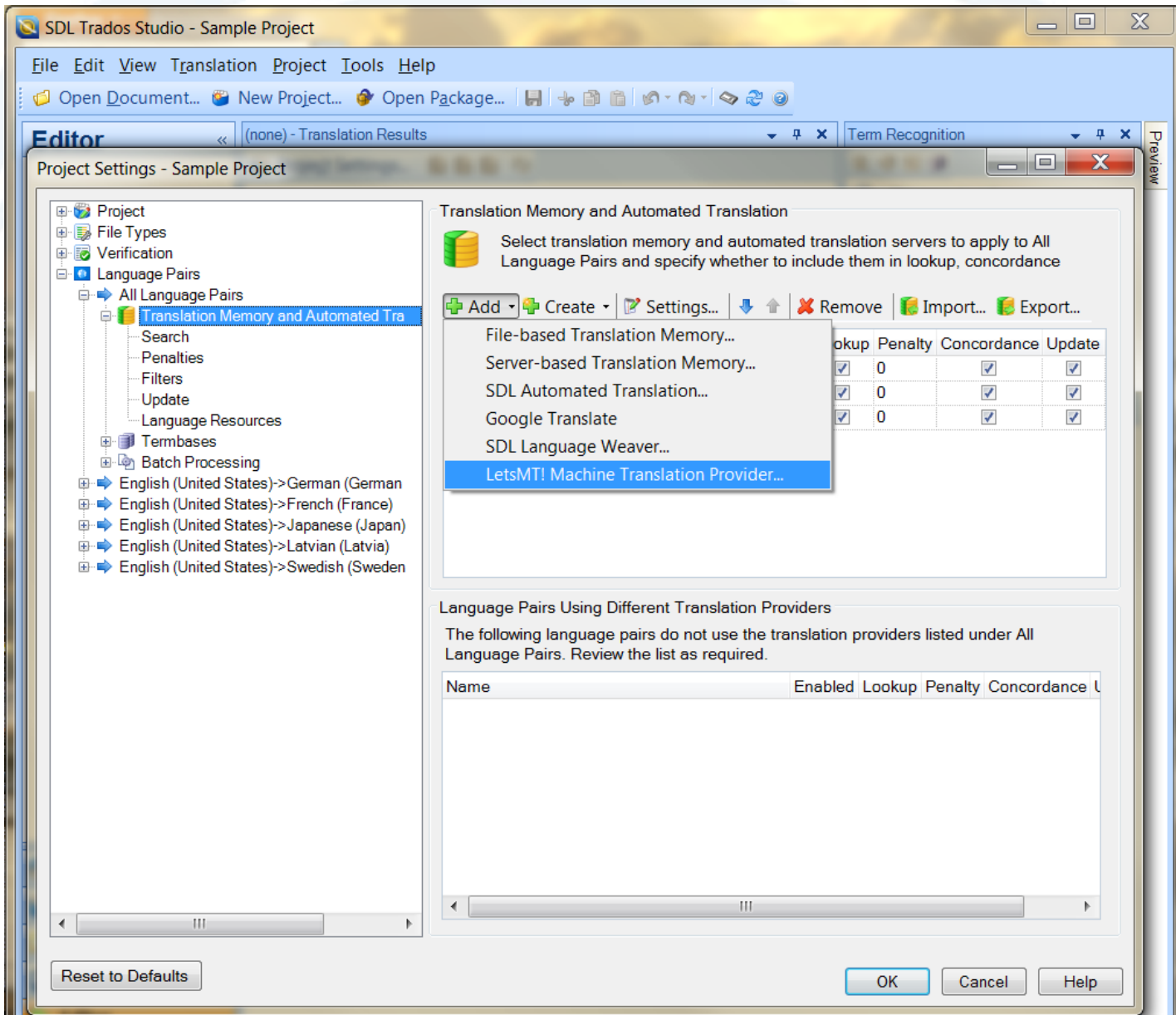
Administrere systemer (og brugere)

LetsMT! Integreret i MemoQ, når næste version af MemoQ releases

Integrere LetsMT! med SDL Trados, mere senere

- Her kun noget om installation af plug-in

# Plug-in for SDL Trados Studio 2011



# Andre faciliteter

## Browser plug-in ([Firefox](#))

- Oversætte direkte fra browser (Firefox)

LetsMT! kan også integreres i web-server, så brugere kan få tilbudt en oversættelse af web-sider

- Wizard kan oversættes til relevant sprog, når der er behov.
- Fx <http://letsmt.newssentiment.eu/>

Press and news

- Press releases
- News
- Press contacts
- Subscribe
- About Nordea Press and news

Blogs

Responsible returns

Categories

- Analysis & Research
- Economics
- Financial reports
- Savings & Investments
- Sustainability
- Awards
- Sponsorship

Photos

Nordea Master brand  
Group Executive

# Nordea best Debt Bank in the Nordic region

2012-04-11 10:35 [News](#)

The magazine Global Finance each year appoints the world's best investment banks in different categories. This year, in the category "Debt banks", Global Finance awarded Nordea as the best debt bank in the Nordic Region. The editors of the magazine, with input from industry experts, used a number of criteria, including market share, number and size of deals, service and distribution network, efforts to address market conditions and market performance of underwritings and market research. Deals announced or completed in 2011 were considered.

Please us this [link](#) to read more about the award

### Tags

[Awards](#), [Global Finance](#)

### Share with others



[Recommend](#) [Sign Up](#) to see what your friends recommend

### Other news



**Exchanging ideas about...**  
2012-04-11 | [Antti Savil](#)

Translation:  
Bladet Global Finance årligt vælger verdens bedste investeringsbanker i forskellige kategorier. Årets bedste innovation i kategorien "Gæld banker, Global Finance udpeget som det bedste gæld hos Nordea i NORDEN. De editors af magasinet med industrien og eksperter inden for input fra en lang række kriterier — markedsandel, inklusive eksponeringen mod flere og større handler, service og rådgivning, strukturering kompetencer, distribution, at adressere nye markedsbetingelser, prisniveau i markedet, efter-underwritings og omdømme i markedet for at vælge de — winners. Handler meddelt afsluttet i 2011 eller i betragtning.  
translated from:  
The magazine Global Finance each year appoints the world's best investment banks in different categories. This year, in the category "Debt banks", Global Finance awarded Nordea as the best debt

### News in other languages

- [Norsk](#) (Norge)
- [Dansk](#) (Danmark)

[Svenska](#) (Sverige)  
[Suomi](#) (Suomi)

[Customer service facts](#)



[All good ideas about Investments](#)  
["All good is an art"](#)



# Prøv letsmt.eu!

## Adgang for interesserede (User Support Group)

- Teste offentlige systemer
- Uploade træningsdata
- Træne og teste systemer
- Vi vil gerne have feedback

# Opsummering

Visionen er:

- Det skal være nemt at finde træningsdata
- Nemt at bruge sine egne TM'er til træning
- Nem adgang til computerkraft, uden at investere
- Ingen indviklet opsætning
- Integration med nogle værktøjer I allerede bruger

Vi håber nogle af jer som en del af "User Group" har lyst til og mulighed for at

- bidrage med data (public eller private)
- træne specifikke systemer og evaluere oversættelses kvaliteten



The research within the project LetsMT! has received funding from the **ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web**, grant agreement n° 250456.

