

Machine Translation
WiSe 2015/2016



Introduction to Machine Translation

Dr. Mariana Neves

October 12th, 2015

Overview

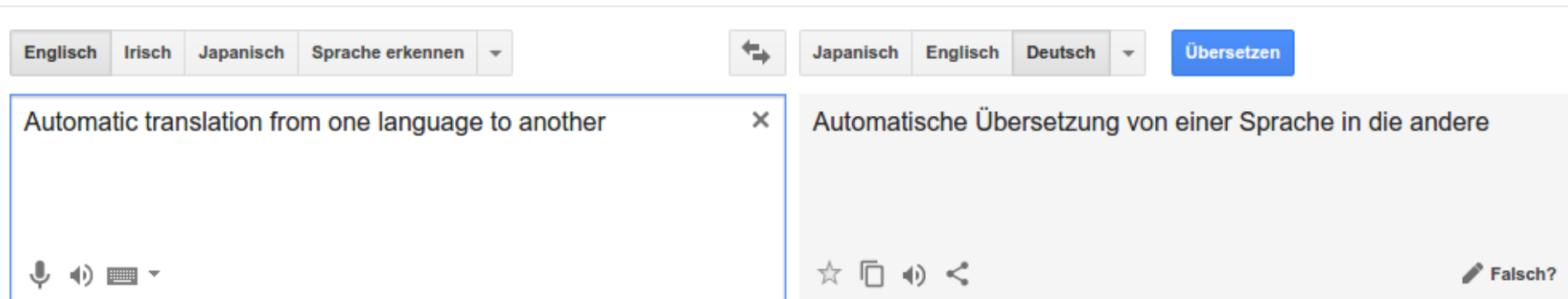
- Introduction
- Applications
- Challenges
- History
- Available resources
- MT course

Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- MT course

Machine translation (MT)

- Automatic translation from one language to another
- Koehn: „Translating between languages is [...] a task for which even humans require special training.“



Machine Translation

C1: DAIYU ALONE ON BED TOP THINK BAOCHAI

E1: *As she* lay there alone Daiyu's thoughts *turned to* Baochai .

C3: CLEAR COLD PENETRATE CURTAIN

E3: *The* coldness penetrated *the* curtains *of her bed* .

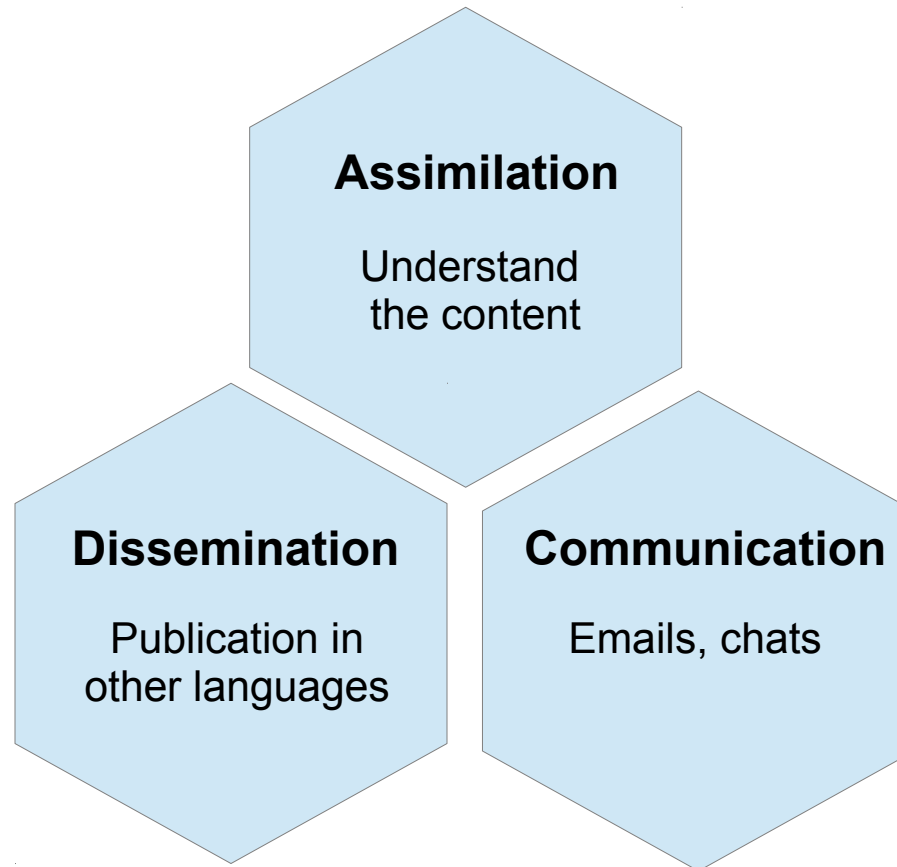
C4: NOT FEELING FALL DOWN TEARS COME

E4: *Almost* without noticing *it she had began to* cry .

Overview

- Introduction
- **Applications**
- Challenges
- History
- Available resources
- MT course

MT applications



FAHQMT

Fully Automatic High Quality Machine Translation

- Limited domains (weather, sport, rail, flight info)
- Controlled vocabulary

Météo: translation of weather reports and warnings

Since 2004, RALI has been investigating how well current SMT approaches deal with a real-world task. We have reconstructed translation systems for dealing with weather bulletins and warnings issued by the [Canadian Meteorological Centre of Environment Canada](#).

#	ED	PER	REP	Source	SMT	REF
1	0	0.000	21	THESE THUNDERSTORMS WILL PRODUCE GUSTY WINDS OF 90 KM / H OR MORE , HAIL STONES OF 2 CM OR MORE , HEAVY RAIN AND FREQUENT LIGHTNING .	CES ORAGES PRODUIRONT DES RAFALES DE 90 KM / H OU PLUS - DE LA GRELE DE 2 CM OU PLUS - DE FORTES PLUIES - ET DE NOMBREUX ECLAIRS .	CES ORAGES PRODUIRONT DES RAFALES DE 90 KM / H OU PLUS - DE LA GRELE DE 2 CM OU PLUS - DE FORTES PLUIES - ET DE NOMBREUX ECLAIRS .
2	0	0.000	38	PERSONS IN THESE REGIONS SHOULD TAKE SAFETY PRECAUTIONS AND LISTEN FOR SUBSEQUENT WARNINGS .	LE PUBLIC DES REGIONS CONCERNEES DEVRAIT PRENDRE LES PRECAUTIONS QUI S IMPOSENT ET SURVEILLER L EMISSION D ALERTES SUBSEQUENTES .	LE PUBLIC DES REGIONS CONCERNEES DEVRAIT PRENDRE LES PRECAUTIONS QUI S IMPOSENT ET SURVEILLER L EMISSION D ALERTES SUBSEQUENTES .
3	0	0.000	5	THIS WARNING IS IN EFFECT FROM 2:20 PM TO 4:50 PM EDT .	CETTE ALERTE EST EN VIGUEUR DE 14H20 A 16H50 HAE .	CETTE ALERTE EST EN VIGUEUR DE 14H20 A 16H50 HAE .
4	0	0.000	16	SEVERE THUNDERSTORMS HAVE WEAKENED OR HAVE MOVED OUT OF THE THESE REGIONS .	LES ORAGES VIOLENTS ONT FAIBLI OU ONT QUITTE CES REGIONS .	LES ORAGES VIOLENTS ONT FAIBLI OU ONT QUITTE CES REGIONS .

(<http://rali.iro.umontreal.ca/rali/?q=en/Meteo>)

Controlled languages - rules

RULE 1:

Write sentences that are shorter than 25 words.

RULE 2:

Write sentences that express only one idea.

RULE 3:

Write the same sentence if you want to express the same content.

RULE 4:

Write sentences that are grammatically complete.

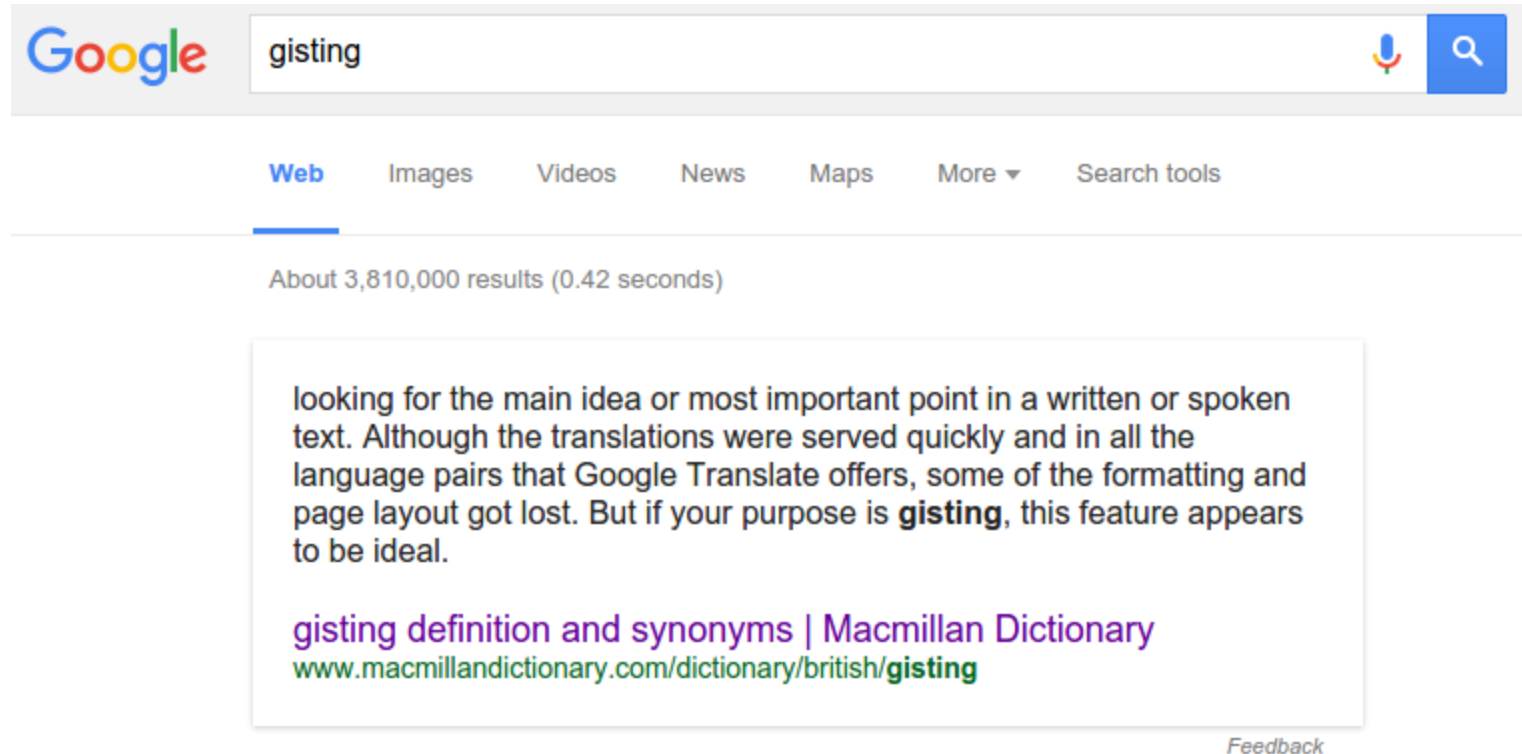
RULE 6:

Write sentences in the active form.

RULE 7:

Write sentences that repeat the noun instead of using a pronoun.

Gisting



The image shows a Google search interface. The search bar contains the word "gisting". Below the search bar, the "Web" tab is selected. The search results show "About 3,810,000 results (0.42 seconds)". A snippet from Macmillan Dictionary is displayed, defining "gisting" as looking for the main idea or most important point in a written or spoken text. The snippet includes a link to the dictionary entry.

Google

gisting

Web Images Videos News Maps More Search tools

About 3,810,000 results (0.42 seconds)

looking for the main idea or most important point in a written or spoken text. Although the translations were served quickly and in all the language pairs that Google Translate offers, some of the formatting and page layout got lost. But if your purpose is **gisting**, this feature appears to be ideal.

[gisting definition and synonyms | Macmillan Dictionary](http://www.macmillandictionary.com/dictionary/british/gisting)
www.macmillandictionary.com/dictionary/british/gisting

Feedback

Gisting

Jobs Property Archives E-Paper Tickets Subscription service



Berliner Morgenpost



Home Berlin Districts Politics Economics Sports From around the world Culture Know Travel Lifestyle Health

SUBSCRIPTION & Apps Specials Service

In the news: Refugees | In western Berlin | Top News via WhatsApp | All Topics

UPDATED REFUGEES

Budapest Station open again - pushing refugees in train



Police have released the train station in Budapest in the morning, the refugees huddled

Most Read Articles

1. TRAINS FROM BUDAPEST
thousands of refugees expected:
Berlin confiscated buildings
2. VOLLSPERRUNG
car transporter overturned -
locked A24 full
3. POLICE AND TRAFFIC
That happened at night to
Thursday in Berlin
4. **UPDATED** REFUGEES
Budapest Station open again -
pushing refugees in train
5. NEW FEATURES
on the U12 rolls a subway called
"Icke"

ANZEIGE

Gisting for intelligence agencies

3PO. And there are other translation projects in the works, such as the recently announced \$5.9 million contract with Raytheon BBN Technologies, to create a real-time English translation of documents, including handwritten notes or images with text on them.

Enter BOLT, which Darpa has asked Congress to fund at \$15 million this year. Once developed, BOLT would act something like C-3PO from the *Star Wars* movies, performing a variety of difficult translation feats for troops in hostile territory.



A soldier translator in Afghanistan

Photograph by Ted Aljibe/AFP/Getty Images.

(<http://info.moravia.com/blog/bid/193094/U-S-defense-projects-may-drive-innovations-in-machine-translation>
http://www.slate.com/articles/technology/future_tense/2012/05/darpa_s_transtac_bolt_and_other_machine_translation_programs_search_for_meaning_.html
<http://www.wired.com/2011/04/militarys-newest-recruit-c-3p0/>)

Gisting for intelligence agencies

- As a first step to select relevant documents from a large collection.
- Interesting documents will then be passed to a human translator

Integration with speech technologies

The screenshot shows the Skype Translator Preview page. The main heading is "Welcome to Skype Translator Preview" with a sub-heading "Now including two additional spoken languages — German & French — and 50 IM languages". A green "Download now" button is visible. Two callout boxes highlight supported languages: one for voice/video calls and one for Instant Messages.

skype Downloads Business Rates Help Sign in Join us

Welcome to Skype Translator Preview

Now including two additional spoken languages — German & French — and 50 IM languages

Download now ↓

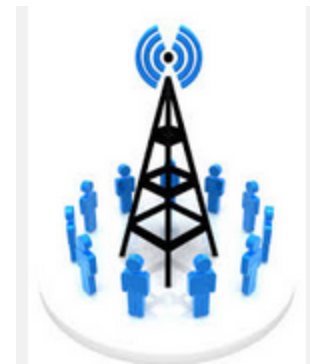
Currently supported for voice/video calls

- Chinese
- English
- French
- German
- Italian
- Spanish

Currently supported for Instant Messages

We currently support 50 languages for Instant Messaging. You can find the [full list here](#).

Integration with speech technologies



Broadcast news speech-to-text translation experiments

Sylvain Raybaud

David Langlois

Kamel Smaili

LORIA - Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cedex
givenname.lastname@loria.fr

Development of SRI's Translation Systems for Broadcast News and Broadcast Conversations

Jing Zheng, Wen Wang, Necip Fazil Ayan

Speech Technology and Research Laboratory, SRI International
{zj,wwang,nfa}@speech.sri.com

A Machine Translation System for Foreign News in Satellite Broadcasting

Teruaki Aizawa, Terumasa Ehara**, Noriyoshi Uratani, Hideki Tanaka,
Naoto Kato, Sumio Nakase*, Norikazu Aruga*, and Takeo Matsuda*

Hand-held devices

Police



military



medical

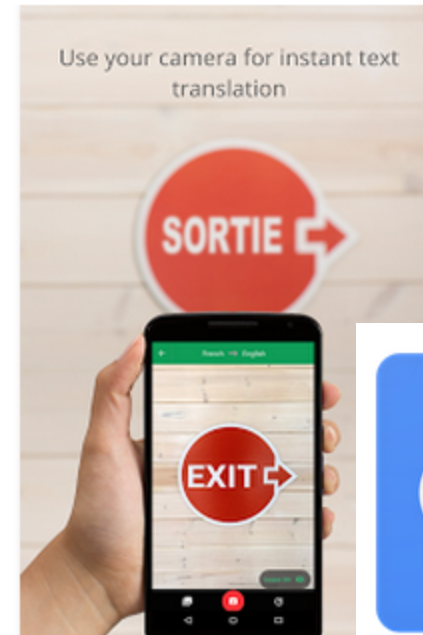


tourism



(<http://www.ectaco.translation.net/>)

Hand-held devices

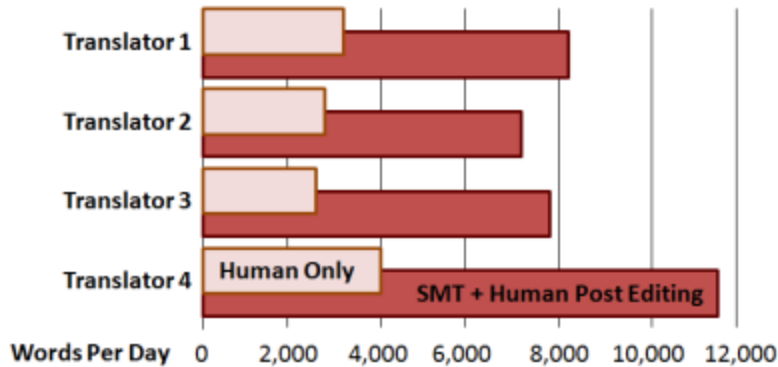
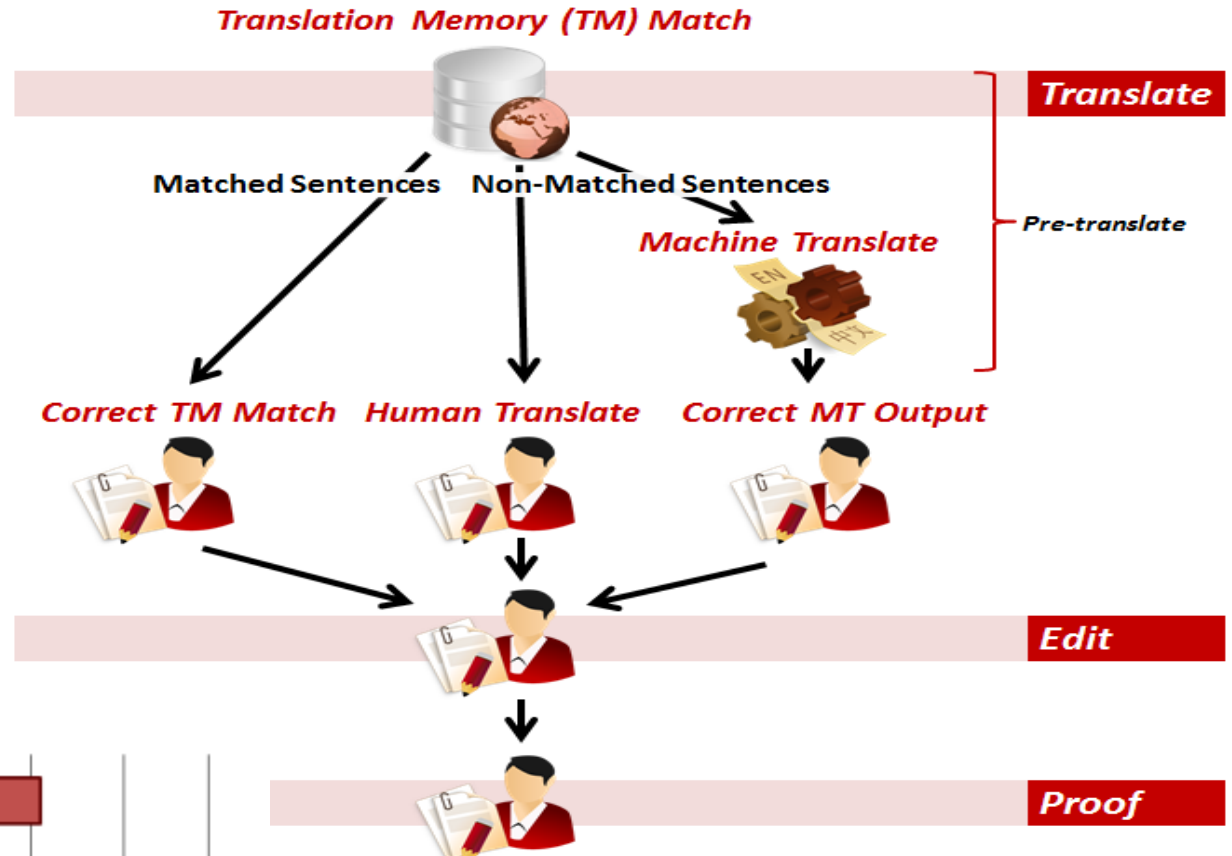


(<http://www.ectaco.translation.net/>)

(<http://www.amazon.com/Bidirectional-Electronic-Dictionary-PhraseBook-Handheld/dp/B001OTMELY>)

(<https://play.google.com/store/apps/details?id=com.google.android.apps.translate&hl=en>)

Tools for translators, Post-editing



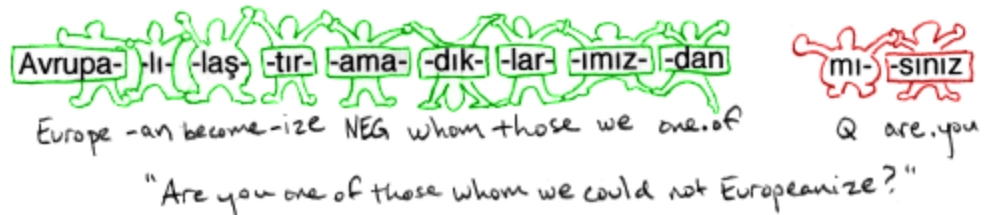
Overview

- Introduction
- Applications
- **Challenges**
- History
- Available resources
- MT course

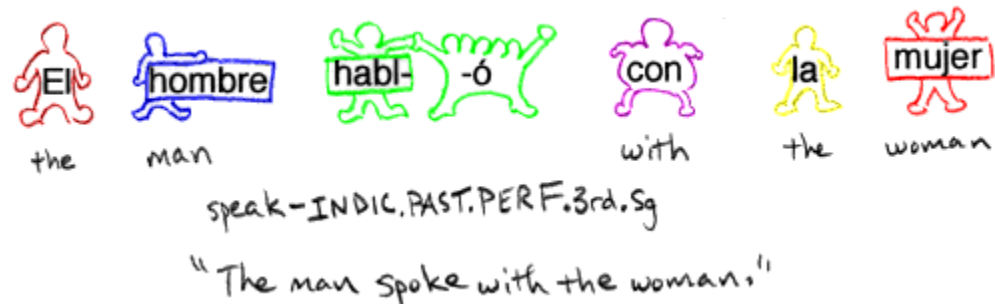
Typology

- Study of cross-linguistic similarities and differences
- Morphology

- Agglutinative
 - Turkish



- Fusion
 - Spanish



Typology

- Syntax: order of verbs (V), subjects (S) and objects (O)

SVO:

(German, French,
English, Mandarin)

She adores listening to music.

SOV:

(Hindi, Japanese)

彼女は音楽を聴いて大好き。

(she music to listening adores)

VSO:

(Irish, Arabic, Biblical Hebrew)

Dúil mhór aici éisteacht le ceol.

(adores she music to listen)

Typology

- Argument structure and linking
 - Head-marking
 - Dependent-marking
 - English: „the man's house“
 - Hungarian: „A férfi házában“ „the man house-his“

Typology

- Verbs and satellite particles (direction, motion, etc.)
- Verb-framed
 - Spanish: „La botella **salió flotando**“ (The bottle **exited floating.**)
 - approach, exit, reach, enter
 - Japanese, Tamil, Romance, Semitic, Mayan

Typology

- Verbs and satellite particles (direction, motion, etc.)
- Satellite-framed
 - English: „The bottle **floated out.**“ (particle)
 - crawl out, float off, jump down, run after
 - English, Swedish, Russia, Indi, Farsi

Typology

- Pronouns omission
 - Pronoun-drop:
 - English: [I] am reading a book.
 - Spanish: Estoy leyendo un libro.

Typology

- Pronouns omission
 - Referential density
 - Cold: more inferential work to recover antecedents
 - Japanese, Chinese
 - Hot: more explicit and easier
 - Spanish

Lexical

- Homonymy
 - wall (**W**and), wall (**M**auer)
- Polysemy
 - to know (knowing a fact) : **w**issen
 - to know (familiarity with a person/location): **k**ennen

Lexical

- Grammar
 - English: „She **likes to** sing“
 - German: „Sie singt **gern.**“
- Lexical gap
 - „A world view, a philosophy of life“ – **Weltanschauung**

Other divergences

- Position of adjectives
 - English: „green witch“
 - Spanish: „bruja verde“ - „witch green“

Other divergences

- Cultural aspects, e.g., calendars and dates
 - British English: **DD/MM/YY**
 - American English: **MM/DD/YY**
 - Japanese: **YYMMDD**

Overview

- Introduction
- Applications
- Challenges
- **History**
- Available resources
- MT course

First references to MT

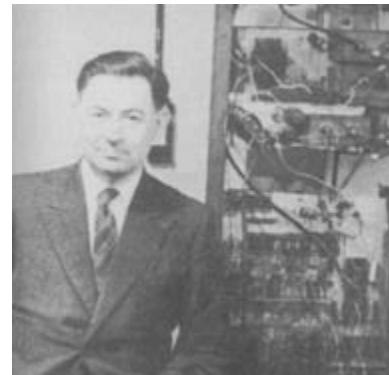
As early as the 17th century by philosophers René Descartes and Gottfried Wilhelm Leibniz



(<http://www.biography.com/people/ren-descartes-37613>
https://en.wikipedia.org/wiki/Gottfried_Wilhelm_Leibniz)

First references to MT

In 1947, Warren Weaver and Andrew Booth suggested that computers could be used to translate natural languages.



(<http://apprendre-math.info/history/photos/Weaver.jpeg>
<http://www.dcs.bbk.ac.uk/about/history/booth.php>)

Post WWII: foreign languages as encrypted English

“One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'”

"Translation" (1955), in W.N. Locke and A.D. Booth (eds.),
Machine Translation of Languages (MIT Press, Cambridge, Mass.).”

Warren Weaver

Scientist

Warren Weaver, PhD was an American scientist, mathematician, and science administrator. He is widely recognized as one of the pioneers of machine translation, and as an important figure in creating support for science in the United States. [Wikipedia](#)



Georgetown-IBM experiment (1954)

- „[...] human translations were subject to political bias and interference“
- Translation of 60 sentences from Russian into English
- Topic: organic chemistry
- System: six grammar rules and 250 words in the vocabulary

Georgetown-IBM experiment (1954)

- Conclusions
 - The problem was solved
 - But semantic disambiguation are impossible to be solved automatically

Russian (Romanized)	English translation
Mi pyeryedayem mislyi posryedstvom ryechyi.	We transmit thoughts by means of speech.
Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radiusu.	Magnitude of angle is determined by the relation of length of arc to radius.
Myezhdunarodnoye ponyimaniye yavlyayetsya vazhnim faktorom v ryeshyenyiyi polyityichyeskix voprosov.	International understanding constitutes an important factor in decision of political questions.

(https://en.wikipedia.org/wiki/Georgetown-IBM_experiment)

ALPAC report (1966)



- Automatic Language Processing Advisory Committee
- Study of reality of MT
- Conclusions:
 - post-editing not cheaper than full translation
 - Little Russian scientific literature worth to be translated
 - No shortage of human translators
 - No advantage in using machine translation
 - Better fund linguistic research for human translation
- Funding for MT stopped in the US as a consequence

History of MT

- 1970s, first commercial systems
 - Météo
 - Systran
 - Logos
 - METAL
 - Trados

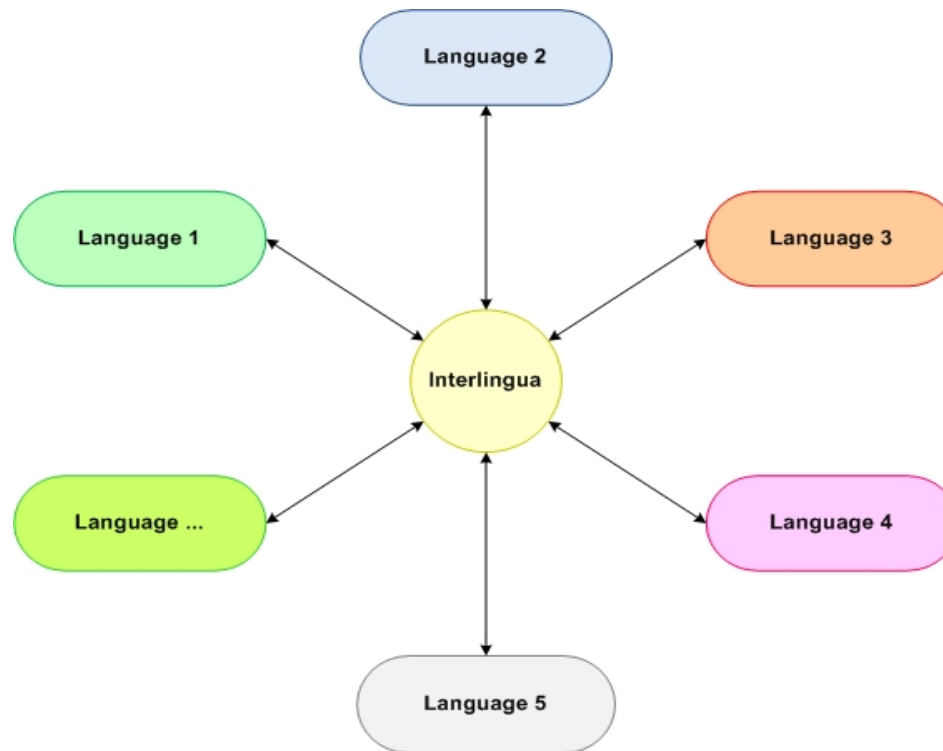
First commercial systems



- 1968: Founded by Dr. Peter Toma
- 1969: US Air Force - scientific and technical documents
Russian/English
- 1975: Commission of European Communities (CEC)
- 1976: CEC – system from English/France
- 1981: CEC – English/French, French/English, English/Italian
- 1986: Xerox – six target languages
- 1985: SYSTRAN PRO for Windows
- 1997: search engine AltaVista's (today Yahoo's) 
- 2006-2007: 

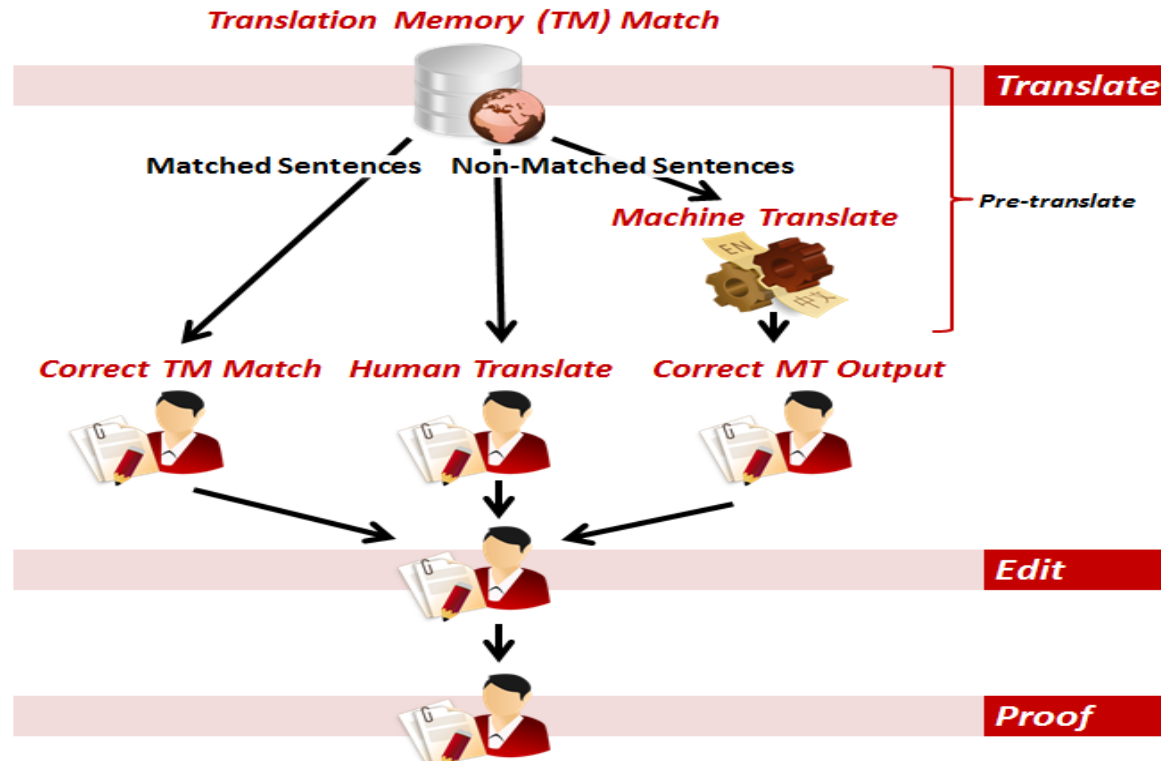
History of MT

- 1980s, 1990s: interlingual systems



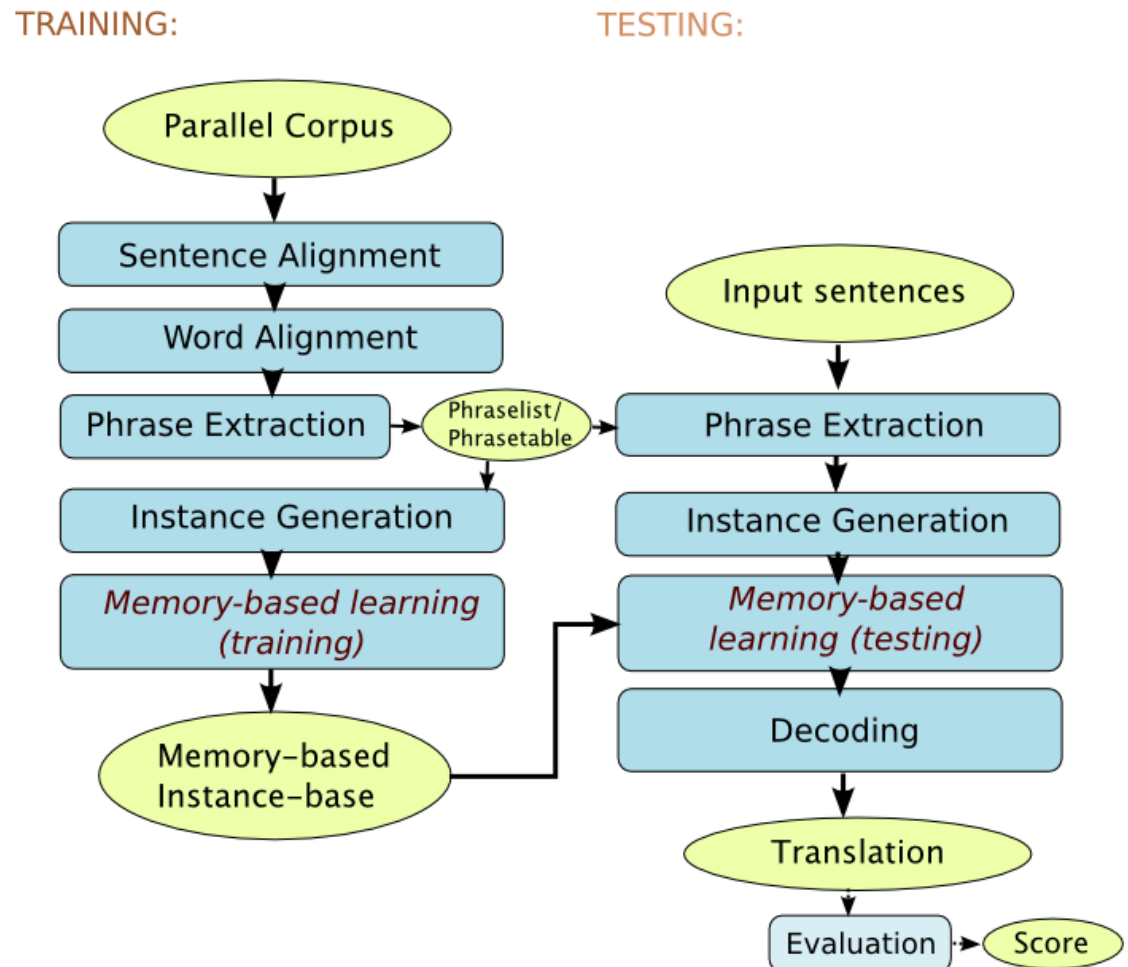
Data-driven methods

- Translation memory



Data-driven methods

- 1980s, Example-based translation



Data-driven methods

- Late 1980, Statistical machine translation

“Most state-of-the-art commercial machine translation systems in use today have been developed using a rules-based approach and require a lot of work by linguists to define vocabularies and grammars. Several research systems, including ours, take a different approach: we feed the computer with billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. We then apply statistical learning techniques to build a translation model.”



Current commercial developers



Overview

- Introduction
- Applications
- History
- Challenges
- **Available resources**
- MT course

Available resources

- Tools
- Parallel corpora

Tools

- MT tools:
 - GIZA++: IBM's word-based models
 - Moses, Thot: phrase-based models
 - SAMT: tree-based models
- MT evaluation tools:
 - BLEU, METEOR

Parallel corpora

- LDC, Gigaword



Arabic Gigaword Fifth Edition

English Gigaword Fifth Edition

(<https://catalog ldc.upenn.edu/LDC2011T11>
<https://catalog ldc.upenn.edu/LDC2011T07>)

Parallel corpora

- Europarl
 - [source release](#) (text files), 1.5 GB
 - [tools](#) (preprocessing tools and sentence aligner only), 8.6 KB
 - [parallel corpus Bulgarian-English](#), 41 MB, 01/2007-11/2011
 - [parallel corpus Czech-English](#), 60 MB, 01/2007-11/2011
 - [parallel corpus Danish-English](#), 179 MB, 04/1996-11/2011
 - [parallel corpus German-English](#), 189 MB, 04/1996-11/2011
 - [parallel corpus Greek-English](#), 145 MB, 04/1996-11/2011
 - [parallel corpus Spanish-English](#), 187 MB, 04/1996-11/2011
 - [parallel corpus Estonian-English](#), 57 MB, 01/2007-11/2011
 - [parallel corpus Finnish-English](#), 179 MB, 01/1997-11/2011
 - [parallel corpus French-English](#), 194 MB, 04/1996-11/2011
 - [parallel corpus Hungarian-English](#), 59 MB, 01/2007-11/2011
 - [parallel corpus Italian-English](#), 188 MB, 04/1996-11/2011
 - [parallel corpus Lithuanian-English](#), 57 MB, 01/2007-11/2011
 - [parallel corpus Latvian-English](#), 57 MB, 01/2007-11/2011
 - [parallel corpus Dutch-English](#), 190 MB, 04/1996-11/2011
 - [parallel corpus Polish-English](#), 59 MB, 01/2007-11/2011
 - [parallel corpus Portuguese-English](#), 189 MB, 04/1996-11/2011
 - [parallel corpus Romanian-English](#), 37 MB, 01/2007-11/2011
 - [parallel corpus Slovak-English](#), 59 MB, 01/2007-11/2011
 - [parallel corpus Slovene-English](#), 54 MB, 01/2007-11/2011
 - [parallel corpus Swedish-English](#), 171 MB, 01/1997-11/2011

Parallel corpora

- Acquis Communautaire

Language ISO Code	N° of Texts	Text body			Signature	Annex	Total N° Words (Text + Signature + Annex)
		Total N° Words	Total N° Characters	Average N° Words	Total N° Words	Total N° Words	
cs	7983	5979261	38479314	749	609441	2100301	8689003
da	7939	6548461	44444011	825	691894	1599456	8839811
de	7914	6576633	47047334	831	571928	1506847	8654608
el	7782	7377316	47715936	948	559487	1628451	9565254
en	7972	7512013	45150120	942	667978	1752545	9932536
es	7809	7964255	48281455	1020	709279	1832745	10506279
et	7944	4925361	38603952	620	439184	1819226	7183771
fi	7735	5134294	43705813	664	565226	1180877	6880397
fr	7862	7812577	45609935	994	673061	1726720	10212358
hu	7489	5391810	40601868	720	539967	1887476	7819253
it	7872	7264126	46792286	923	707467	1704221	9675814
lt	7966	5386359	39936370	676	625365	1948354	7960078
lv	7980	5656335	39290110	709	461736	2011426	8129497
mt	7639	7230538	43919981	947	505324	2288013	10023875
nl	7882	7339465	47699598	931	712255	1710041	9761761
pl	7968	5974605	43160945	750	668248	2070687	8713540
pt	7848	7851904	47225710	1001	648180	1838833	10338917
ro	5792	5122354	33681450	884	402929	4047393	9572676
sk	5278	3911895	26077956	741	413511	1381471	5706877
sl	7984	5989322	37844883	750	573052	2153138	8715512
sv	7731	6472717	42990411	837	560188	1424887	8457792
Average	7,636	6,353,410	42,340,925	831	585,947	1,886,338	8,825,695

Parallel corpora



Scientific Electronic Library Online

Anais da Sociedade Entomológica do Brasil

versão impressa ISSN 0301-8059

Resumo

VIEIRA, Marineide R. e CHIAVEGATO, Luiz G.. Biologia de *Polyphagotarsonemus latus* (Banks) (Acari: Tarsonemidae) em limão siciliano (*Citrus limon* Burm). An. Soc. Entomol. Bras. [online]. 1999, vol.28, n.1, pp. 27-33. ISSN 0301-8059. <http://dx.doi.org/10.1590/S0301-80591999000100002>.

No estudo da biologia de *Polyphagotarsonemus latus* em limão Siciliano, foram utilizados potes plásticos circulares com capacidade de 250 ml, contendo areia esterilizada como suporte para dois frutos novos com aproximadamente 2,0 cm de diâmetro. O ensaio foi conduzido a $27,1 \pm 0,5^{\circ}\text{C}$, umidade relativa de $67,6 \pm 1,3\%$ e fotofase contínua. O período de ovo a adulto durou $3,7 \pm 0,1$ dias para fêmeas e $3,6 \pm 0,1$ dias para machos, com sobrevivência de 100%. Após um período de pré-oviposição de $1,0 \pm 0,2$ dias, as fêmeas depositaram $5,6 \pm 0,5$ ovos por dia durante $10,5 \pm 0,9$ dias, totalizando $58,9 \pm 6,7$ ovos por fêmea. A longevidade foi de $13,4 \pm 1,0$ dias para fêmeas e $12,0 \pm 2,4$ dias para machos. A razão intrínseca de aumento (rm) foi de 0,359, a razão finita de aumento (l) de 1,43 indivíduos por fêmea por dia, o tempo médio de uma geração (T) de 10,34 dias e a taxa líquida de reprodução (Ro) de 41,0.

Palavras-chave : Ácaro branco; desenvolvimento biológico; tabela de vida de fertilidade; taxa líquida de reprodução.

Anais da Sociedade Entomológica do Brasil

versão impressa ISSN 0301-8059

Resumo

VIEIRA, Marineide R. e CHIAVEGATO, Luiz G.. Biology of *Polyphagotarsonemus latus* (Banks) (Acari: Tarsonemidae) on lemon (*Citrus limon* Burm). An. Soc. Entomol. Bras. [online]. 1999, vol.28, n.1, pp. 27-33. ISSN 0301-8059. <http://dx.doi.org/10.1590/S0301-80591999000100002>.

In the study of the biology of *Polyphagotarsonemus latus* (Banks) on lemon var. Siciliano (*Citrus limon* Burm) round plastic pots (250 ml) containing sterilized sand were used as support for two 2cm-diameter new fruits. The assay was carried out at $27.5 \pm 0.5^{\circ}\text{C}$, relative humidity of $67.6 \pm 1.3\%$ and constant photophase. The duration of immature phases was 3.7 ± 0.1 days for females and 3.6 ± 0.1 days for males, with 100% survival. After a pre-oviposition period of 1.0 ± 0.2 days, the females deposited 5.6 ± 0.5 eggs per day during 10.5 ± 0.9 days, i.e., 58.9 ± 6.7 eggs per female. The longevity was 13.4 ± 1.0 days for females and 12.0 ± 2.4 days for males. The intrinsic rate of increase (rm) was 0.359, finite rate of increase (l) 1.43 individual per female per day, mean generation time (T) 10.34 days and net reproductive rate (Ro) 41.0.

Palavras-chave : Broad mite; biological development; life table of fertility; net reproductive rate.

Evaluation campaigns

NIST Open Machine Translation 2015 Evaluation (OpenMT15)

Highlights

- Evaluation on informal data genres (SMS/Chat, Conversational Telephone Speech) for Arabic-to-English and Chinese-to-English
- Inclusion of audio input track
- Explore common MT measurement techniques on these informal data genres

(<http://www.nist.gov/itl/iad/mig/openmt15.cfm>)

Evaluation campaigns

IWSLT 2015, International Workshop on Spoken Language Translation

- Automatic speech recognition (ASR), i.e. the conversion of a speech signal into a transcript,
- Machine translation (MT), i.e. the translation of a polished transcript into another language,
- Spoken language translation (SLT), i.e. the conversion and translation of a speech signal into a transcript in another language.

Evaluation campaigns

EMNLP 2015 TENTH WORKSHOP ON STATISTICAL MACHINE TRANSLATION

- English-German and German-English
- English-French and French-English
- English-Finnish and Finnish-English **NEW**
- English-Czech and Czech-English
- English-Russian and Russian-English



**HPI at the
WMT'16!!!**

Overview

- Introduction
- Applications
- Challenges
- History
- Available resources
- **MT course**

MT course – what to expect from me

- Overview of MT methods
 - Specially statistical machine translation
- Arrange dataset and additional resources for the exercises
- Be available by email and in the office (Mo 16:45-17:30)

MT course – what I expect from you

- Presence and participation in the lecture (not controlled)
- Hand-in the exercises
- Present the exercises
- Take part in the exam

Lectures

(Program is subject to change)

Week	Date	Topic
1	October 12, 2015	Introduction to Machine Translation
2	October 19, 2015	Words, sentences and corpora
3	October 26, 2015	Classical MT
4	November 2, 2015	Probability theory, Words (students)
5	November 9, 2015	Word-based models
6	November 16, 2015	(no lecture)
7	November 23, 2015	Word-based models (students)
8	November 30, 2015	Phrase-based models
9	December 7, 2015	**Guest talk: Dr. Feiyu Xu (DFKI, Yocoy)**
10	December 14, 2015	Phrase-based models (students)
11	January 4, 2016	Decoding
12	January 11, 2016	Language models
13	January 18, 2016	Evaluation
14	January 25, 2016	tbd
15	February 1, 2016	Final exam (room tbd, 15:15)

Guest talk



Feiyu Xu
徐飞玉



DFKI GmbH
Language Technology Lab

Dr. Feiyu Xu is Senior Researcher and Project Leader in the [Language Technology Lab of DFKI](#).

She also is co-founder of [Yocoy Technologies GmbH](#), a 2007 spin-off from DFKI. Yocoy is developing next generation mobile language and travel guides such as [yochina](#) and [i-You](#).

Since 2004, Dr. Xu is vice-director of the [Joint Research Laboratory](#) for Language Technology of Shanghai Jiao Tong University and Saarland University.

She studied technical translation at Tongji University in Shanghai with a waiver of the national admission exam in year 1987, a high honor for outstanding students. She then studied computational linguistics at Saarland University from 1992 to 1998 and graduated by receiving a Diplom (MSc) with distinction for a thesis on "[Semantic Underspecification and Ellipsis Resolution](#)". She joined DFKI in 1998. In 2007 she obtained her PhD in the area of information extraction for a thesis on "[Bootstrapping Relation Extraction from Semantic Seeds](#)".

She has extensive experience with building multilingual and monolingual information systems and leading ICT projects. Her research areas include information extraction, information mining, semantic web technologies, question answering, opinion mining, dialogue systems, language learning and mobile applications of language technologies. She has been involved in more than ten national, EU and international projects.

Guest talk



yocoy

Yocoy can help you find the best wording. The app both translates and speaks

Exercises

- Still to be decided
 - Many small tasks, every two weeks
- Teams of 2-3 students
- Based on SAP HANA database
 - Queries
 - Stored procedures
- Dataset will be provided (DE/EN)

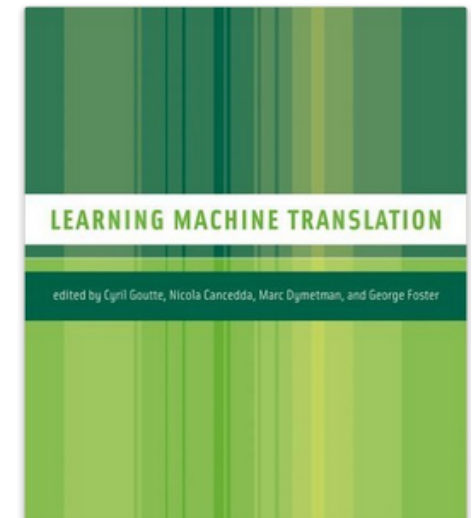
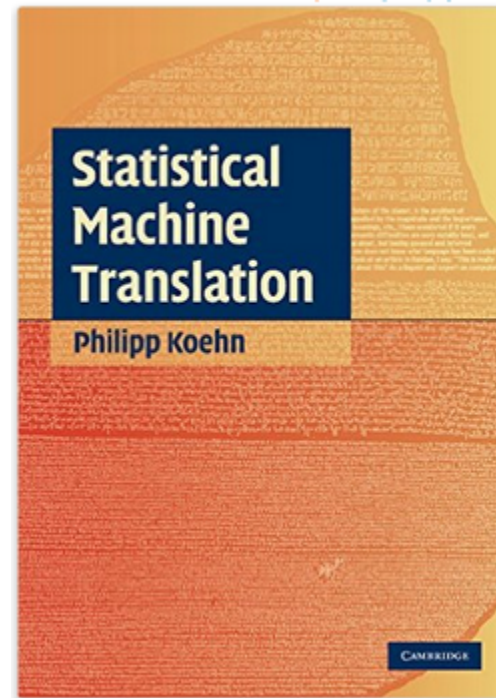
Grading

- All exercises need to be handed
- One exam on February 2nd, 2016

Course books

- Statistical Machine Translation
 - Philipp Koehn

- Learning Machine Translation
 - Edited by Cyril Goutte, Nicola Cancedda, Marc Dymetman



Course books

- Speech and Language Processing
 - Daniel Jurafsky and James H. Martin
 - Chapter 25

