Machine Translation
WiSe 2016/2017

Introduction to Machine Translation

Dr. Mariana Neves

October 17th, 2016

Overview

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

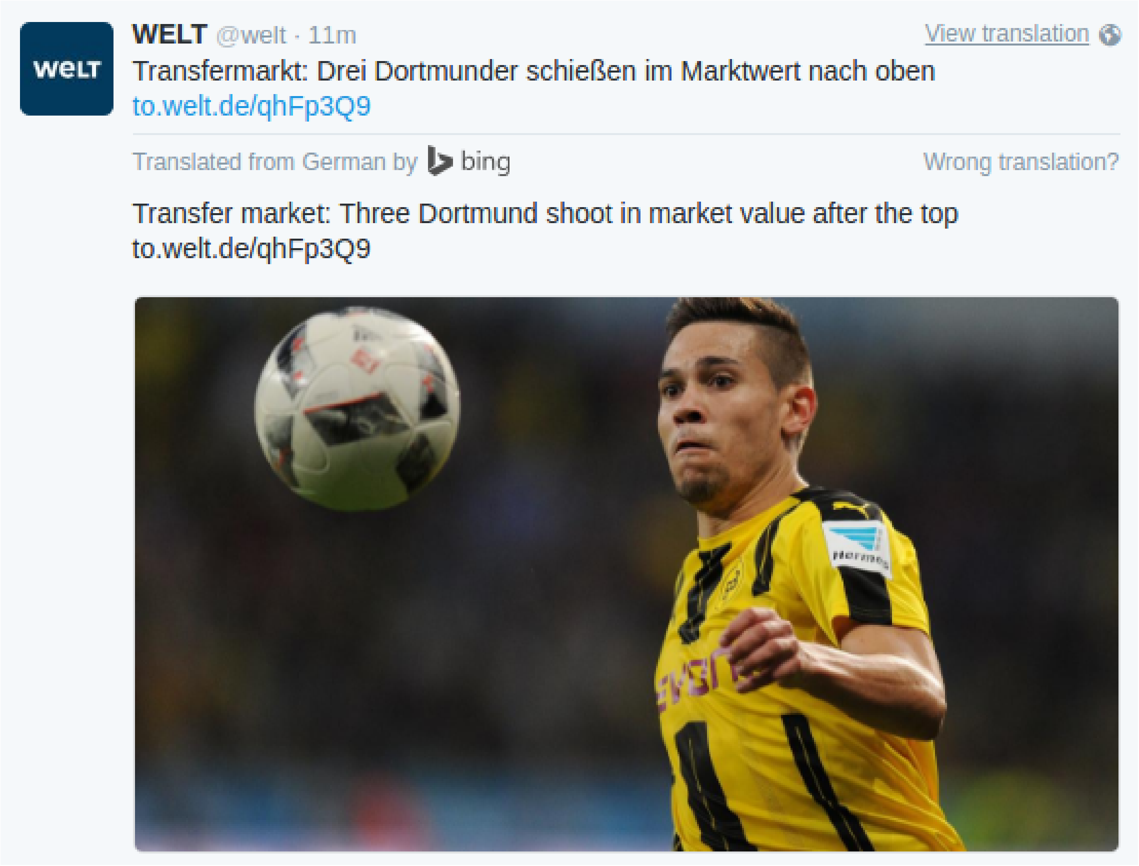
Overview

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



WELT @welt · 11m

View translation

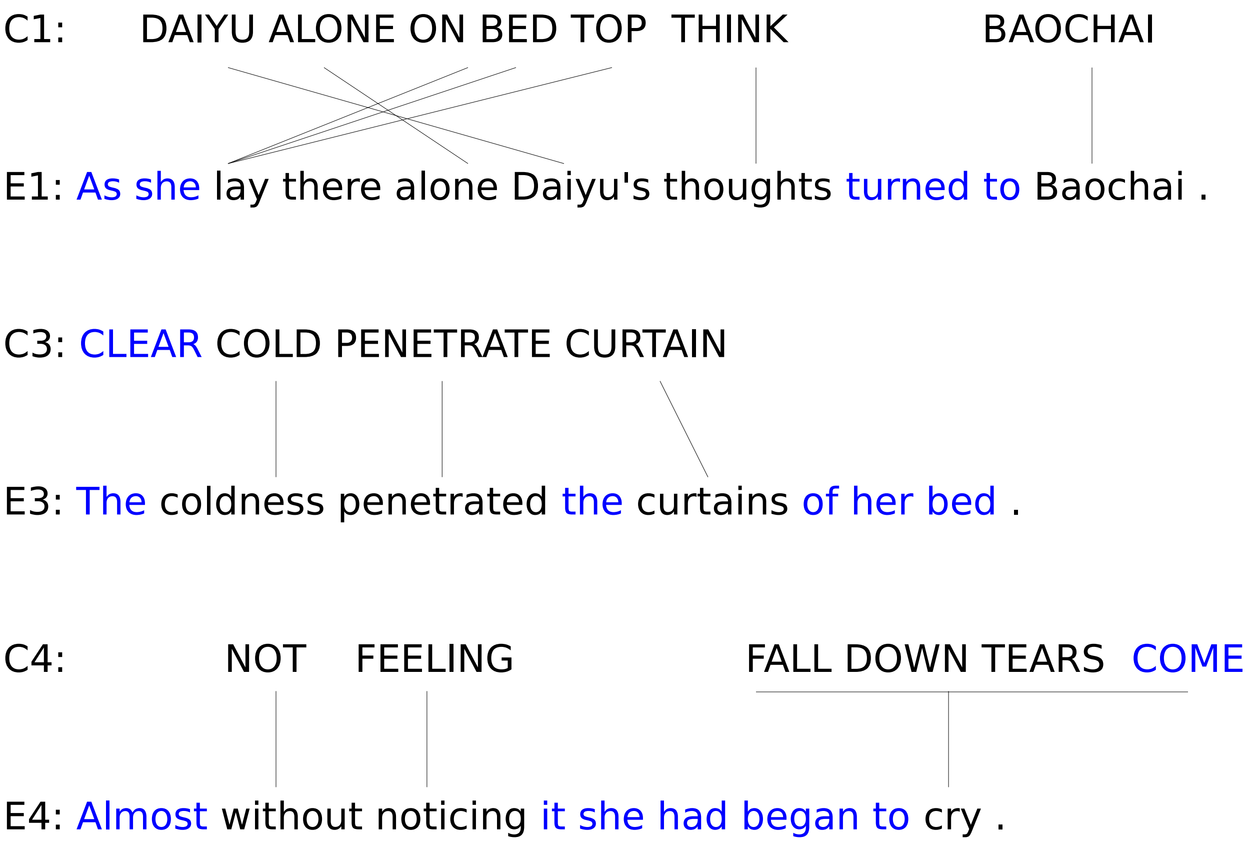
Transfermarkt: Drei Dortmunder schießen im Marktwert nach oben to.welt.de/qhFp3Q9

Translated from German by bing

Wrong translation?

Transfer market: Three Dortmund shoot in market value after the top to.welt.de/qhFp3Q9

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

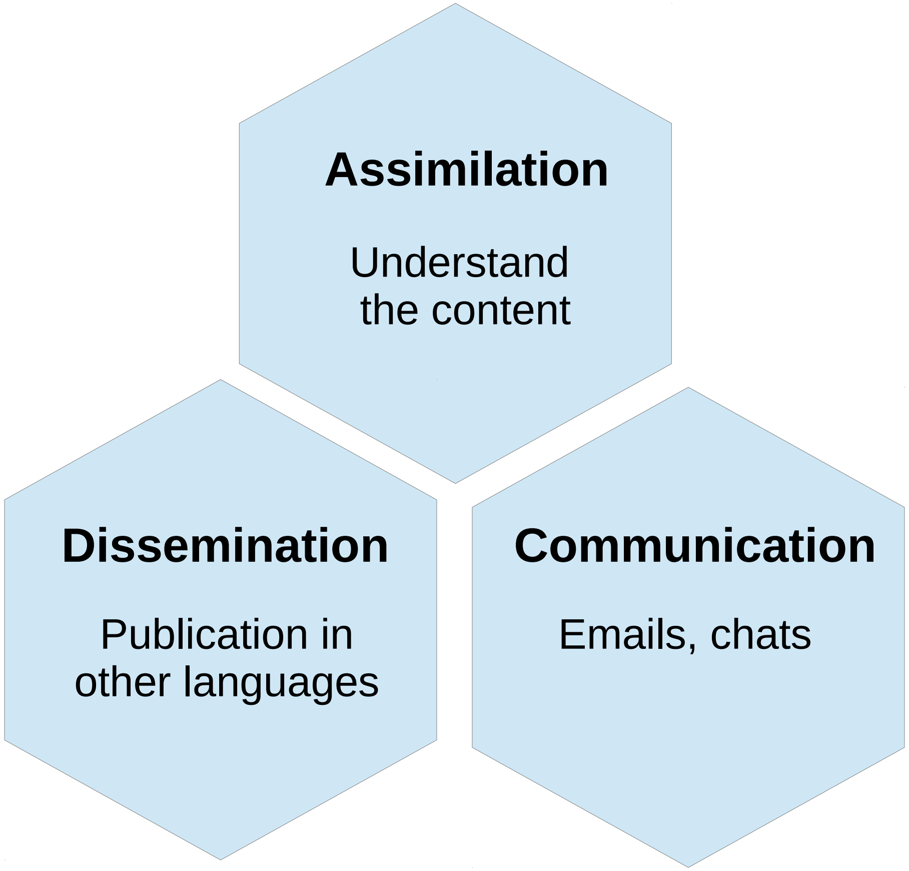
E4: Almost without noticing it she had began to cry .

[From Jurafski and Martin 2009]

Overview

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

MT applications



Assimilation

Understand
the content

Dissemination

Publication in
other languages

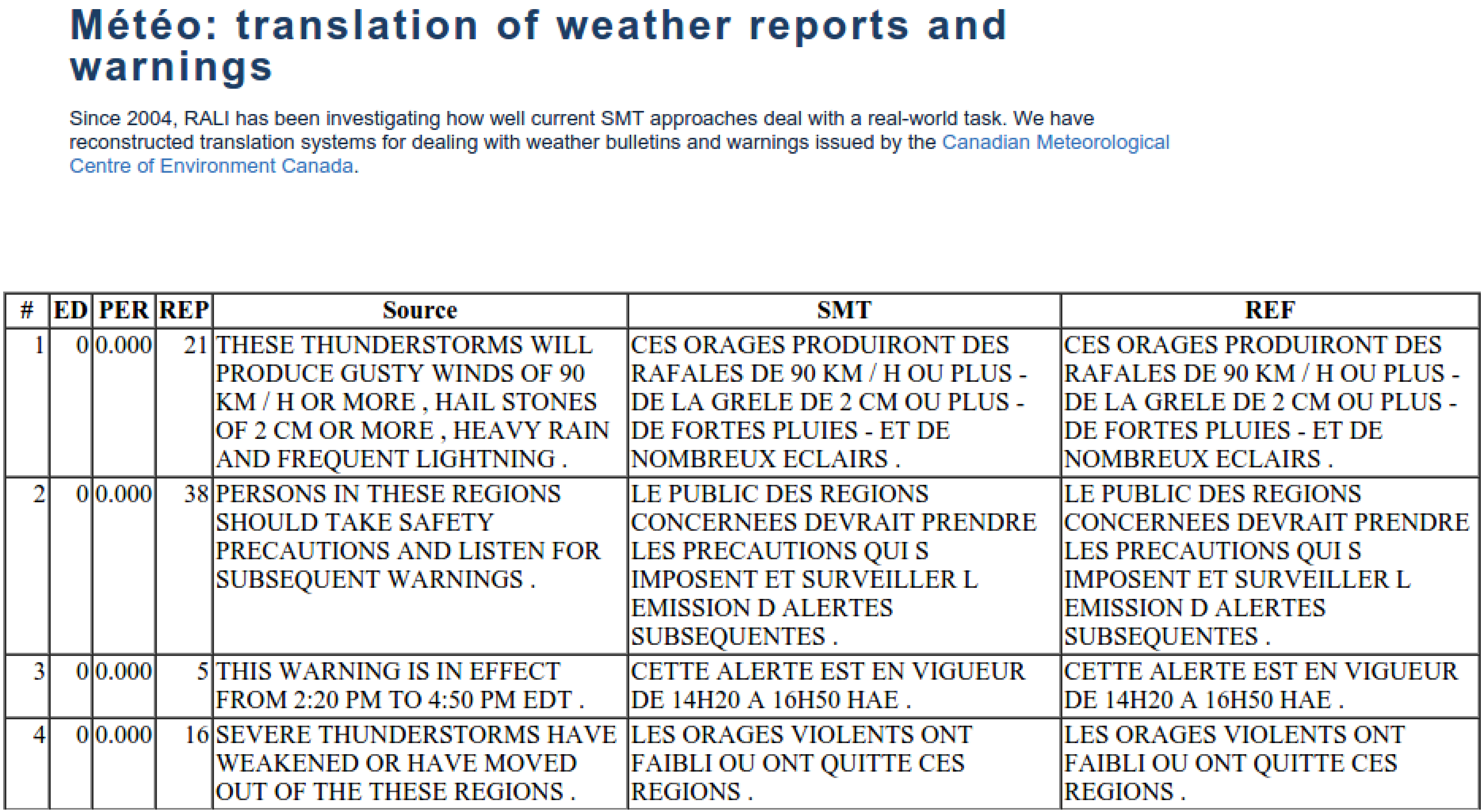
Communication

Emails, chats

[Koehn 2010]

Fully Automatic High Quality Machine Translation (FAHQMT)

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



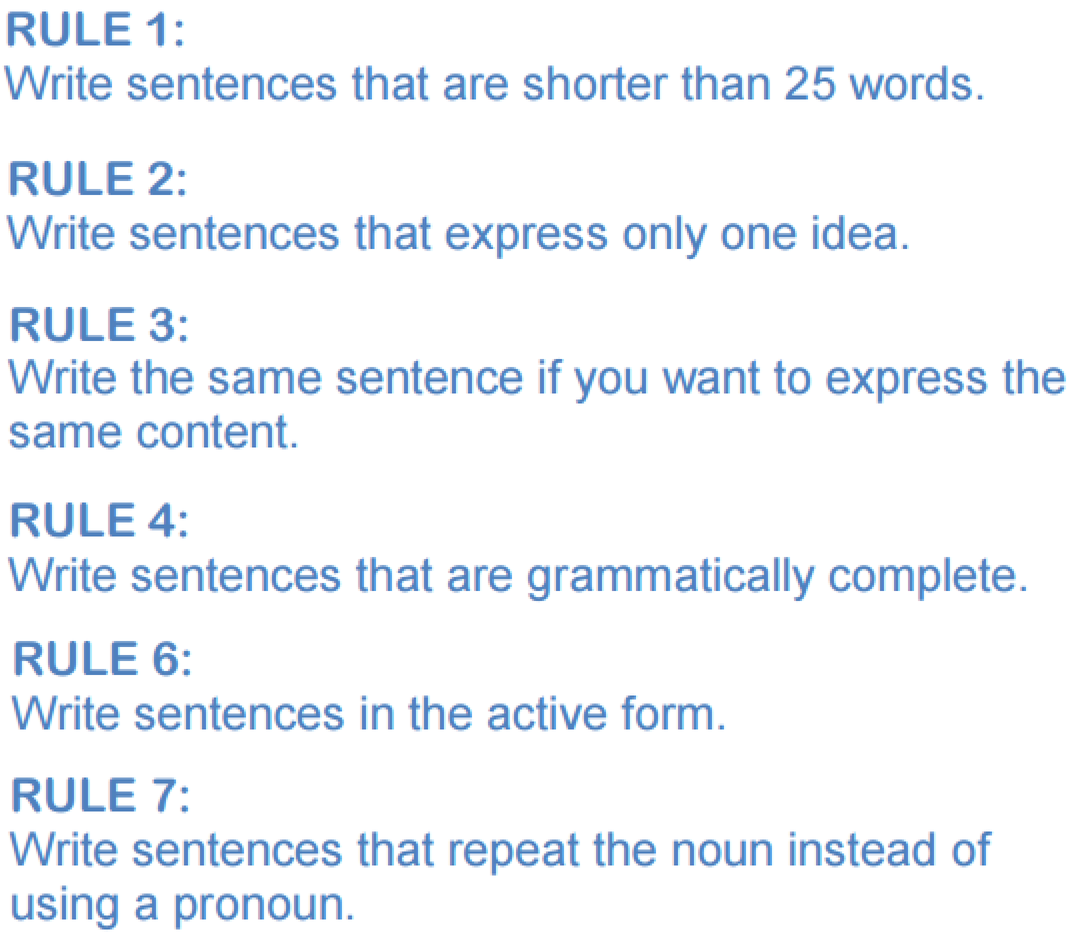
Météo: translation of weather reports and warnings

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.

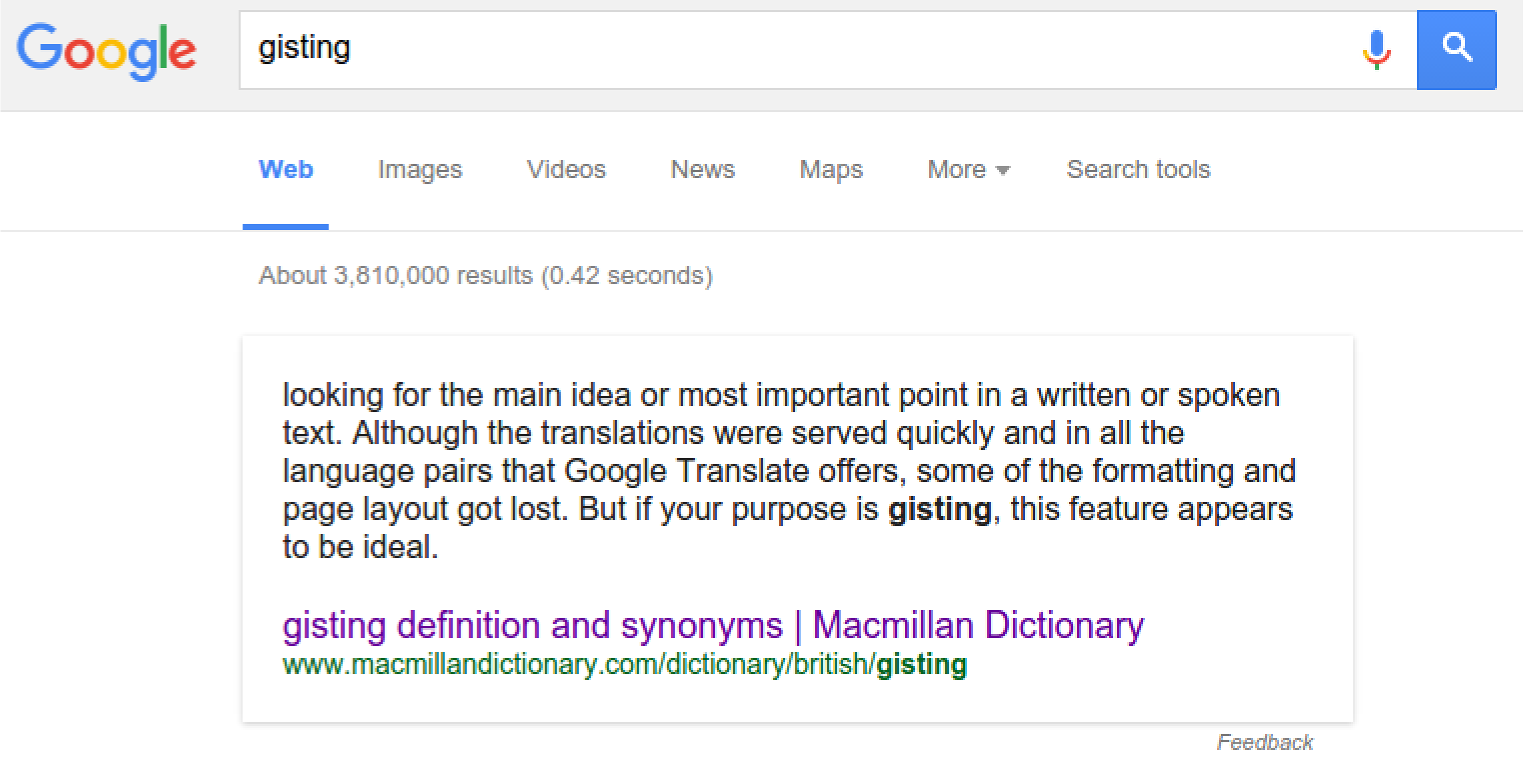
RULE 4:
Write sentences that are grammatically complete.

RULE 6:
Write sentences in the active form.

RULE 7:
Write sentences that repeat the noun instead of using a pronoun.

(http://works.bepress.com/cgi/viewcontent.cgi?article=1126&context=uwe_muegge)

Gisting

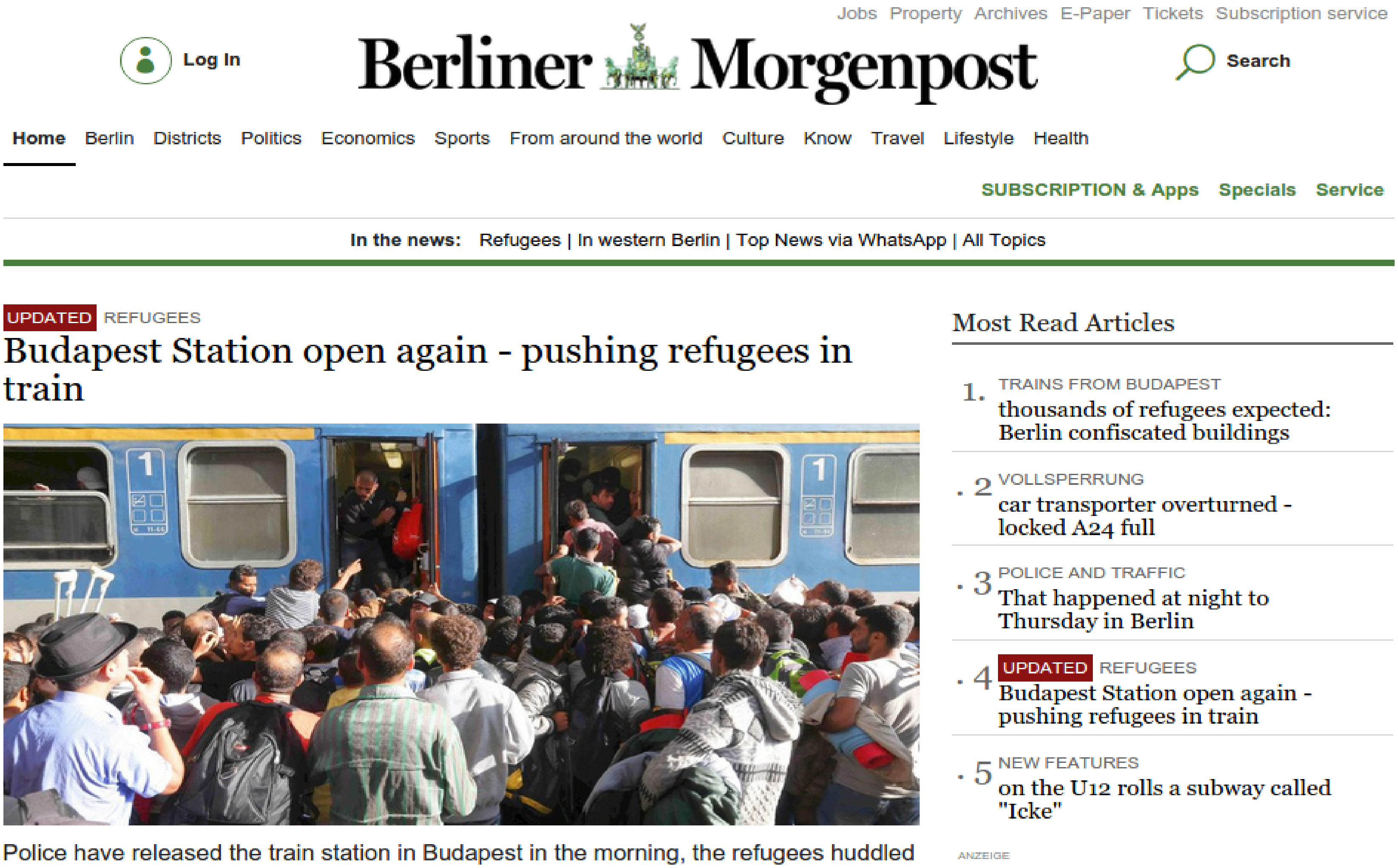


About 3,810,000 results (0.42 seconds)

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
www.macmillandictionary.com/dictionary/british/**gisting**

Gisting



Jobs Property Archives E-Paper Tickets Subscription service

Log In

Berliner Morgenpost

Search

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

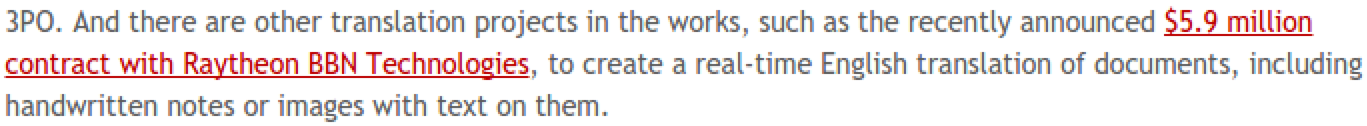
. 4 UPDATED REFUGEES
Budapest Station open again - pushing refugees in train

. 5 NEW FEATURES
on the U12 rolls a subway called "Icke"

ANZEIGE

(http://www.morgenpost.de/)

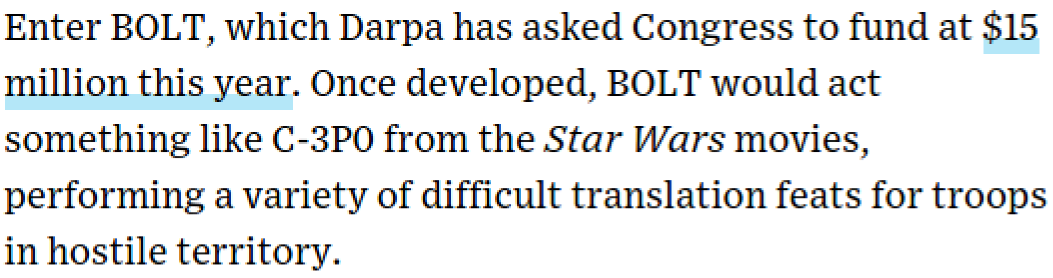
Gisting for intelligence agencies



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.

A soldier translator in Afghanistan

Photograph by Ted Aljibe/AFP/Getty Images.



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.

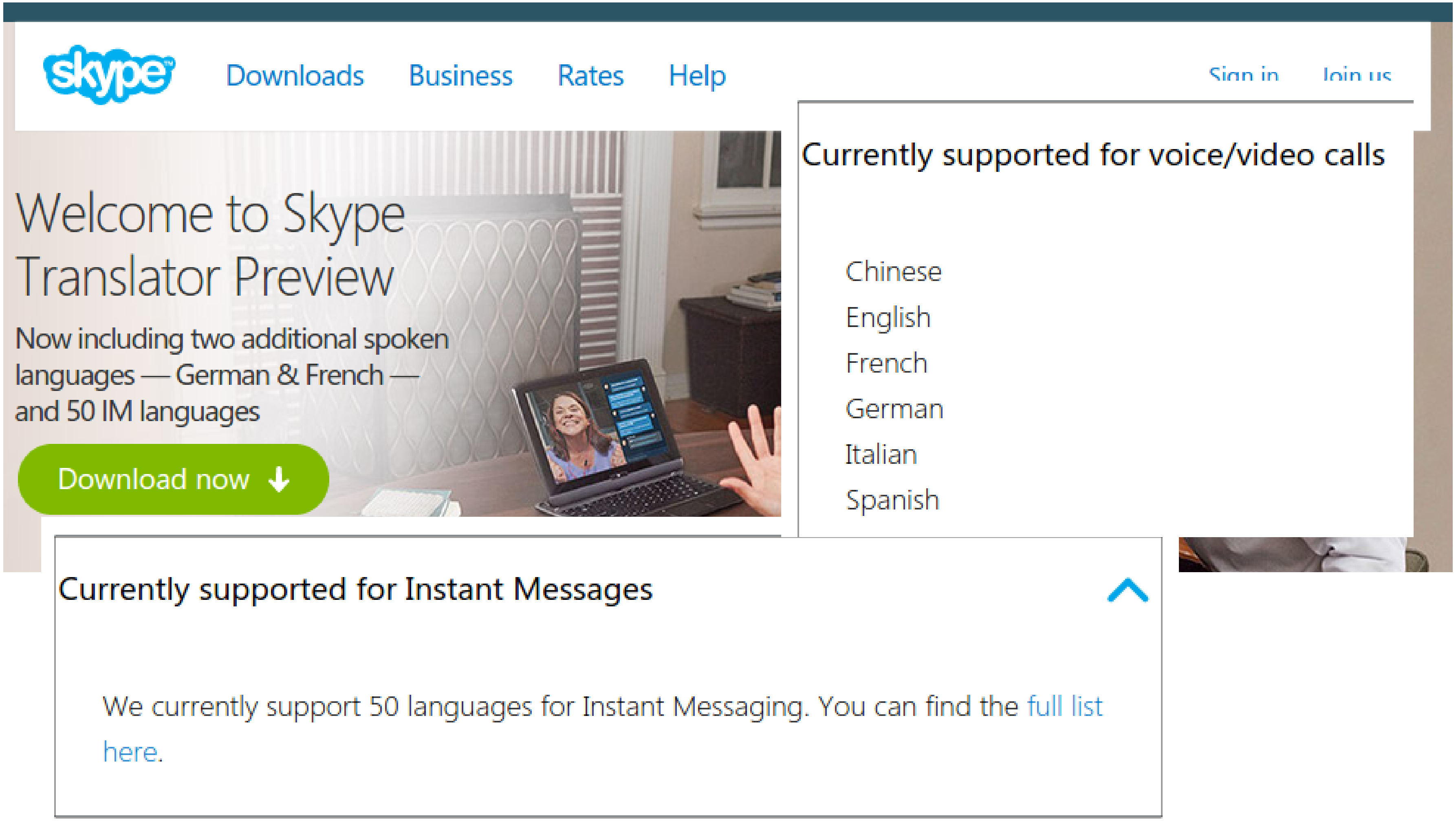
(http://info.moravia.com/blog/bid/193094/U-S-defense-projects-may-drive-innovations-in-machine-translation
http://www.slate.com/articles/technology/future_tense/2012/05/darpa_s_transtac_bolt_and_other_machine_translation_programs_search_for_meaning_.html
http://www.wired.com/2011/04/militarys-newest-recruit-c-3p0/)

Gisting for intelligence agencies

- As a first step, select relevant documents from a large collection.

- Interesting documents will then be passed to a human translator

Integration with speech technologies



Welcome to Skype Translator Preview

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

Currently supported for voice/video calls

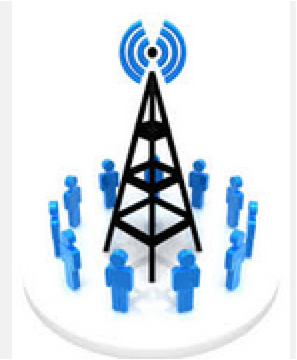
- Chinese
- English
- French
- German
- Italian
- Spanish

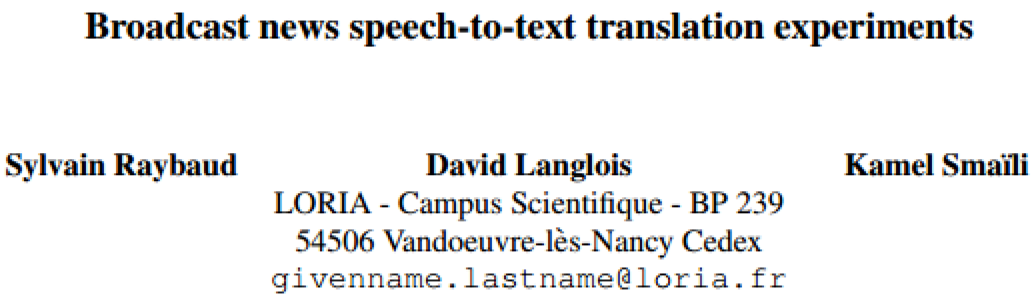
Currently supported for Instant Messages

We currently support 50 languages for Instant Messaging. You can find the full list here.

(http://www.skype.com/en/translator-preview/)

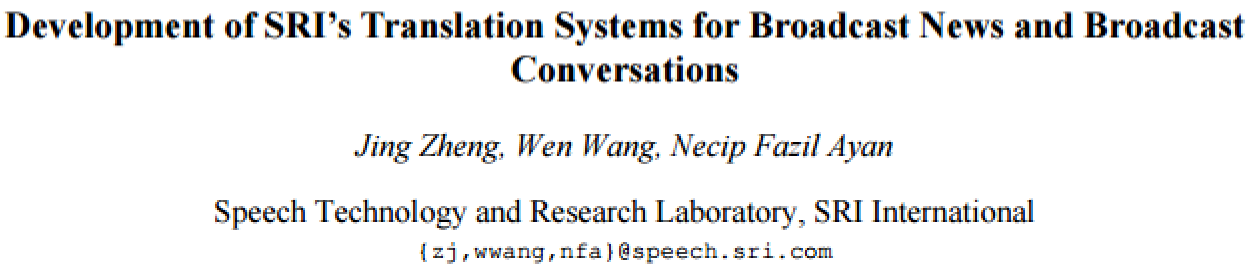
Integration with speech technologies





Broadcast news speech-to-text translation experiments

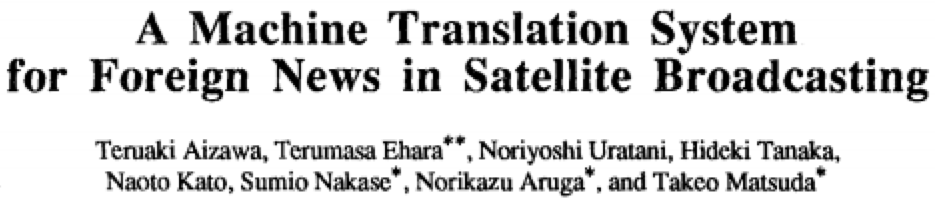
Sylvain Raybaud **David Langlois** **Kamel Smaïli**
LORIA - Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cedex
givenname.lastname@loria.fr



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*

(http://www.itproportal.com/2012/06/26/how-to-broadcasting-your-business-presentation-anywhere-in-the-world/)

Hand-held devices

police

military

medical

tourism

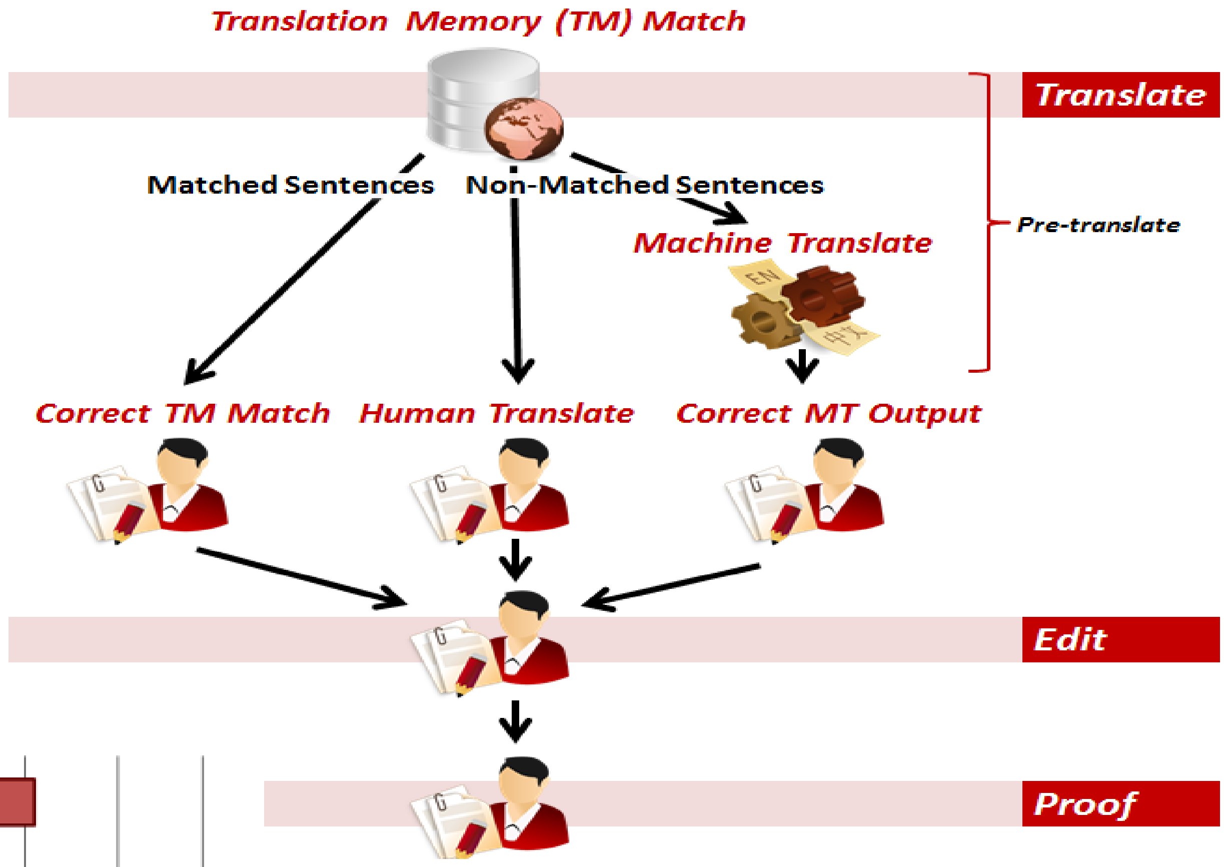
(http://www.ectaco.translation.net/)

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



Translation Memory (TM) Match

Matched Sentences

Non-Matched Sentences

Machine Translate

Translate

Pre-translate

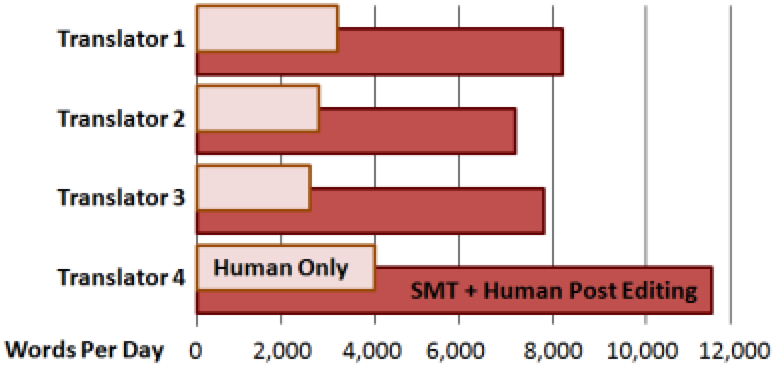
Correct TM Match

Human Translate

Correct MT Output

Edit

Proof



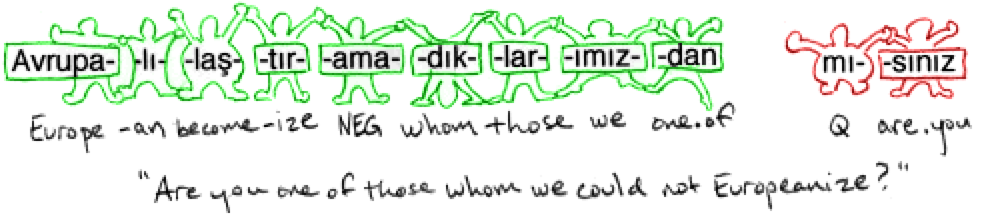
(http://www.languagestudio.com/LanguageStudioDesktop.aspx
http://www.asiaonline.net/EN/MachineTranslation/default.aspx?QID=21)

Overview

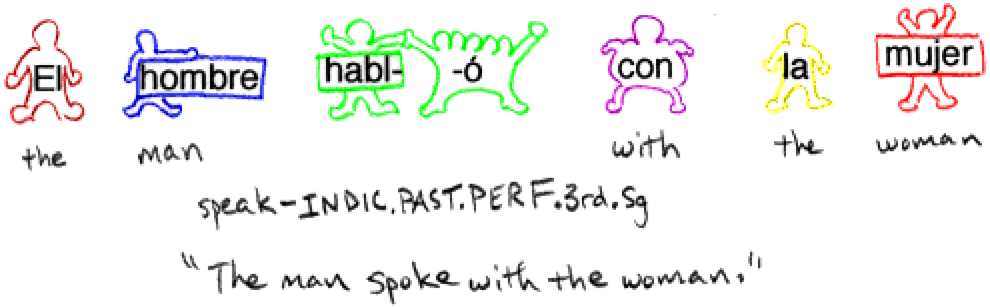
- Introduction
- Applications
- Challenges
- History
- Available resources
- MT paradigms
- MT course

Typology

- Study of cross-linguistic similarities and differences
- Morphology
  - Agglutinative
    - Turkish


Avrupa- -lı- -laş- -tır- -ama- -dık- -lar- -ımız- -dan
mı- -sınız
Europe -an become -ize NEG whom those we one.of
Q are.you
"Are you one of those whom we could not Europeanize?"

  - Fusion
    - Spanish


El hombre habl- -ó con la mujer
the man
with the woman
speak-INDIC.PAST.PERF.3rd.Sg
"The man spoke with the woman."

(http://allthingslinguistic.com/post/50939757945/morphological-typology-illustrations-from)

Typology

- Syntax: e.g., order of verb (V), subject (S) and object (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:
    - „the man's house“ (English)
  - Dependent-marking:
    - „A férfi házában“ „the man house-his“ (Hungarian)

Typology

- Verbs and satellite particles (direction, motion, etc.)

- Verb-framed:
  - Spanish: „La botella salió flotando“ (The bottle exited floating.)

- Satellite-framed
  - English: „the bottle floated out“

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 (Wand), wall (Mauer)

- Polysemy
  - to know (knowing a fact) : wissen
  - to know (familiarity with a person/location): kennen

Lexical

- Grammar
  - English: „She likes to sing“
  - German: „Sie singt gern.“

- Lexical gap
  - „A world view, a philosophy of life“ – Weltanschauung

(http://abbysroad.tumblr.com/post/12947835861/an-incomplete-list-of-english-lexical-gaps)

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 paradigms
- 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.

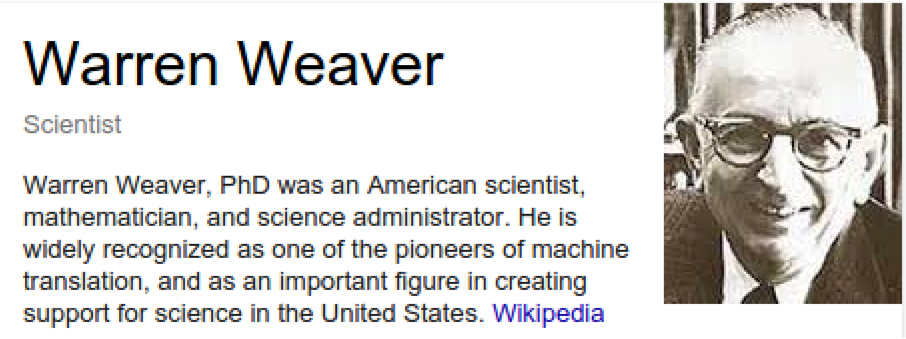
(http://apprendre-math.info/history/photos/Weaver.jpeg
http://www.dcs.bbk.ac.uk/about/history/booth.php)

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

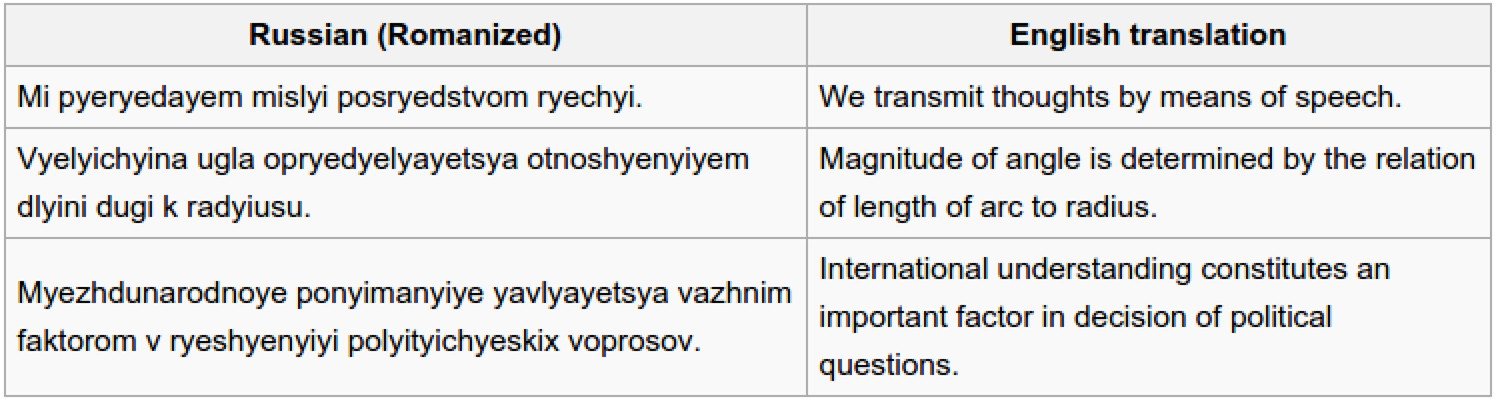
Georgetown-IBM experiment (1954)

- „[…] 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

(Report: http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf
http://www.thelinguafile.com/2013/10/the-georgetown-ibm-experiment-rise-of.html#.Vehb1t93nq5)

Georgetown-IBM experiment (1954)

- Conclusions
  - The problem was solved
  - But semantic disambigution 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 radyiusu.	Magnitude of angle is determined by the relation of length of arc to radius.
Myezhdunarodnoye ponyimanyiye 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

(http://www.hutchinsweb.me.uk/MTNI-14-1996.pdf)

History of MT

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

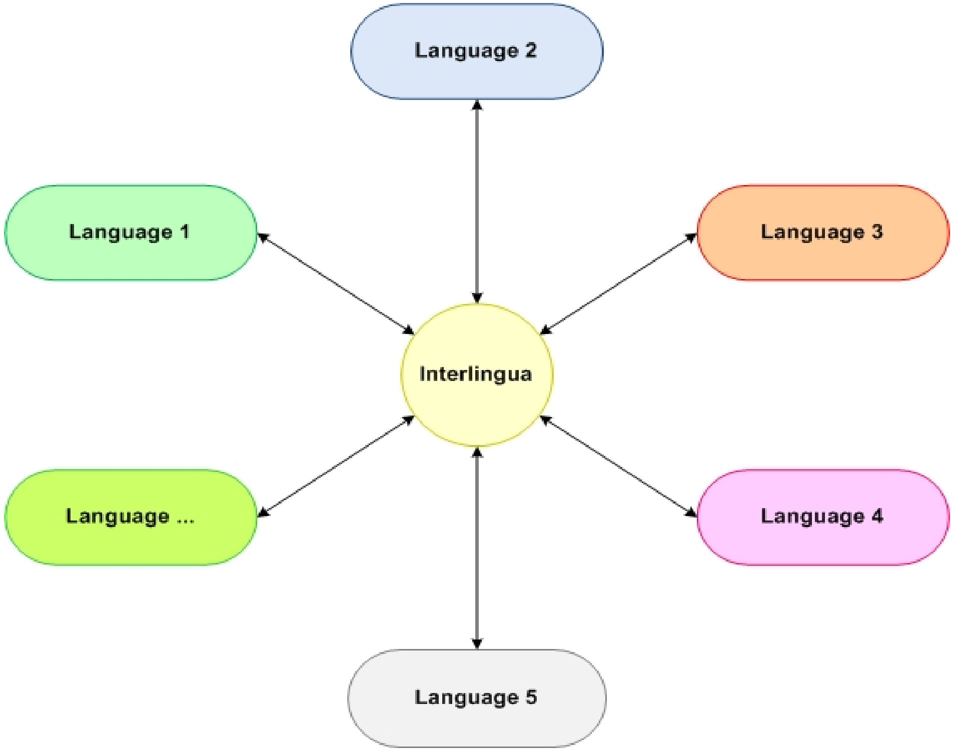
First commercial systems SYSTRAN

- 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) BabelFish
- 2006-2007: Google Translate

(http://www.thelinguafile.com/2013/11/systran-brief-history-of-machine.html#.VehiWd93nq4)

History of MT

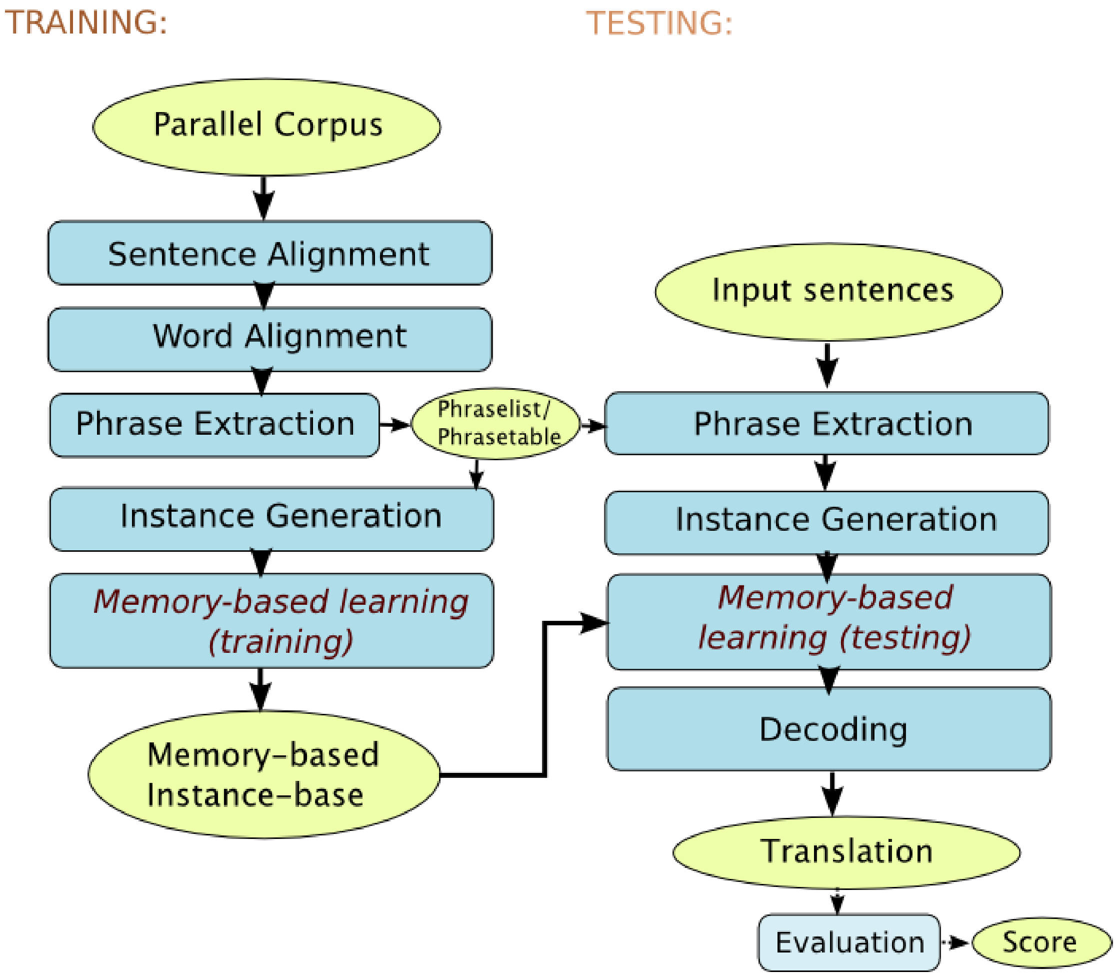
- 1980s, 1990s: interlingual systems



(http://www.dictionarybarn.com/img/Interlingual-Machine-Translation.jpg)

Data-driven methods

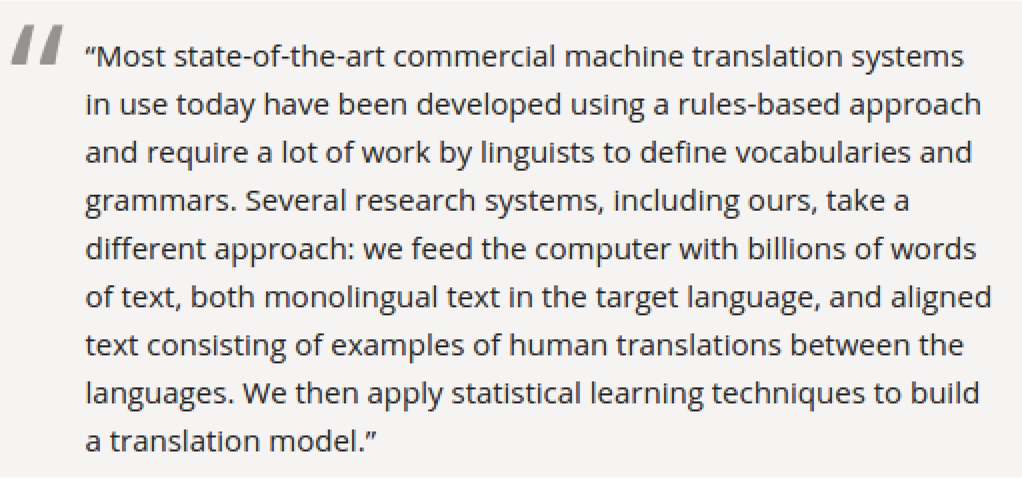
- 1980s, Example-based translation



(http://ilk.uvt.nl/mbmt/pbmbmt/)

Data-driven methods

- Late 1980, Statistical machine translation



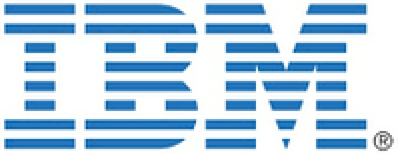
> “Most state-of-the-art commercial machine translation systems in use today have been developed using a rules-based approach and require a lot of work by linguists to define vocabularies and grammars. Several research systems, including ours, take a different approach: we feed the computer with billions 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.”

Google Translate

Break through language barriers.

(http://www.themarysue.com/how-does-google-translate-work/)

Current commercial developers



Overview

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

Available resources

- Tools

- Parallel corpora

Tools

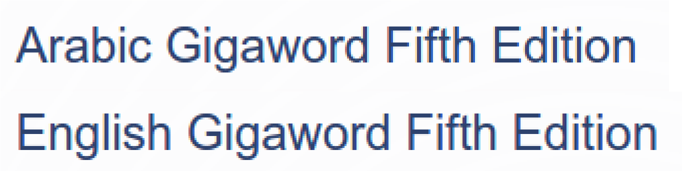
- Natural language processing (NLP) tools:
  - Tokenization, parsing, named-entity recognition

- MT tools:
  - GIZA++: IBM's word-based models
  - Moses, Thot: phrase-based models

- MT evaluation tools:
  - BLEU, METEOR

Parallel corpora

- LDC, Gigaword

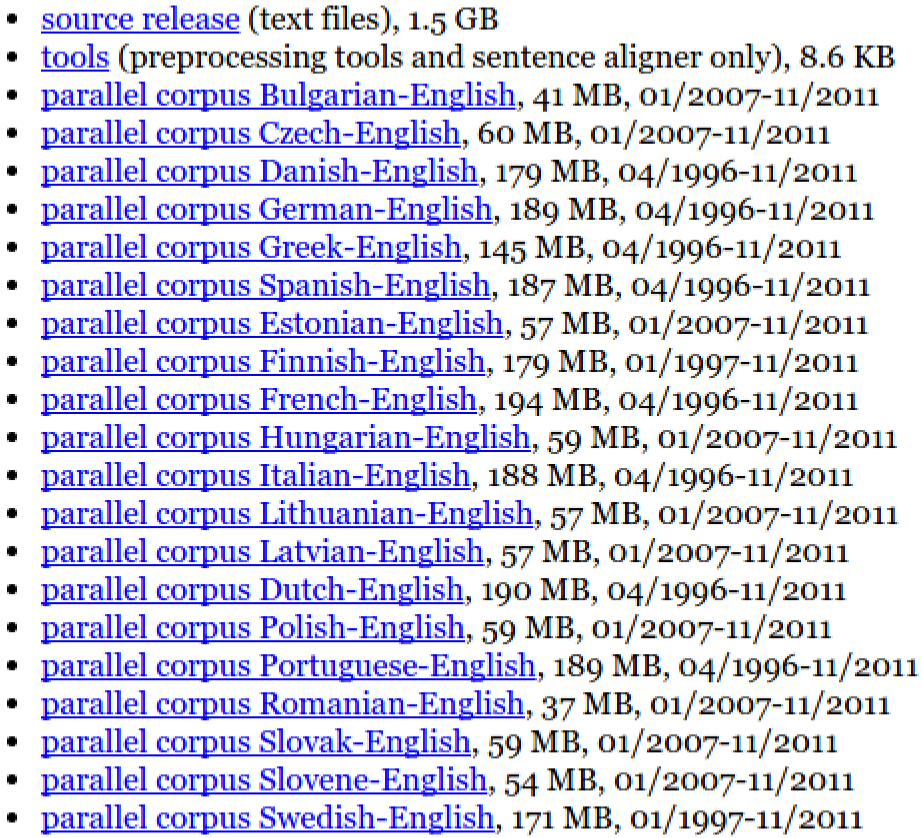

Arabic Gigaword Fifth Edition

English Gigaword Fifth Edition

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

Parallel corpora

- Europarl

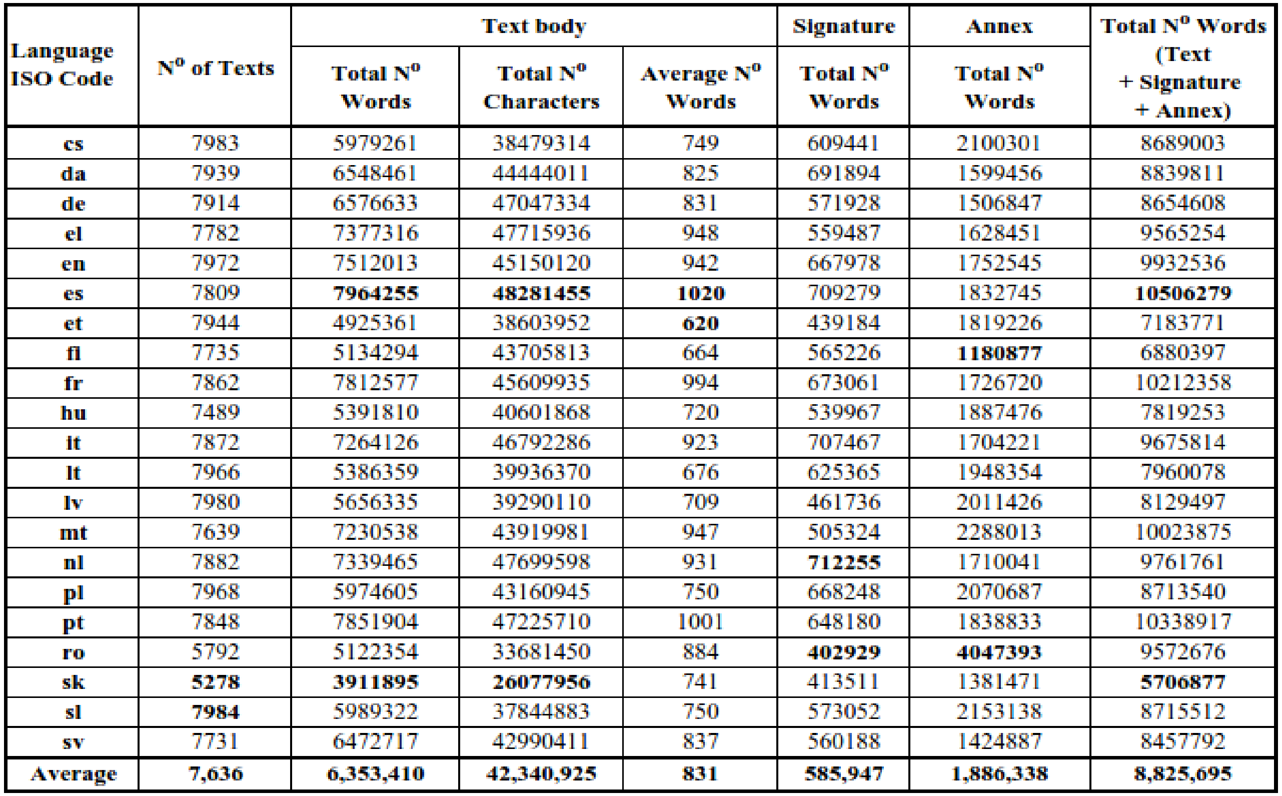


- source release (text files), 1.5 GB
- tools (preprocessing tools and sentence aligner only), 8.6 KB
- parallel corpus Bulgarian-English, 41 MB, 01/2007-11/2011
- parallel corpus Czech-English, 60 MB, 01/2007-11/2011
- parallel corpus Danish-English, 179 MB, 04/1996-11/2011
- parallel corpus German-English, 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

(http://www.statmt.org/europarl/)

Parallel corpora

- Acquis Communautaire

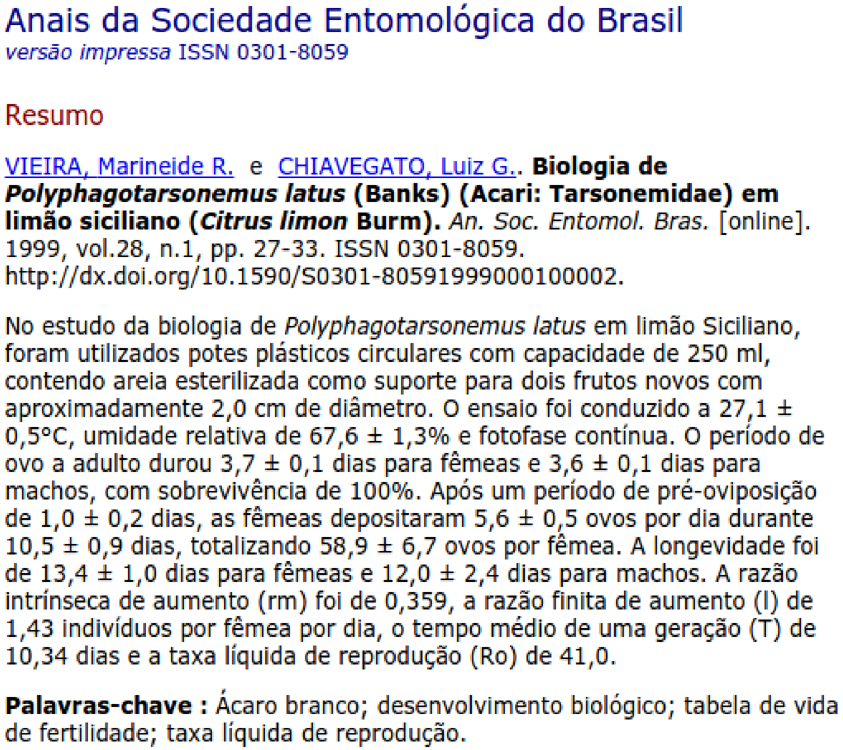


Language ISO Code	Nº of Texts	Text body			Signature	Annex	Total Nº Words (Text + Signature + Annex)
		Total Nº Words	Total Nº Characters	Average Nº Words	Total Nº Words	Total Nº Words	
cs	7983	5979261	38479314	749	609441	2100301	8689003
da	7939	6548461	44444011	825	691894	1599456	8839811
de	7914	6576633	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**

(http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html)

Parallel corpora

Scientific Electronic Library Online



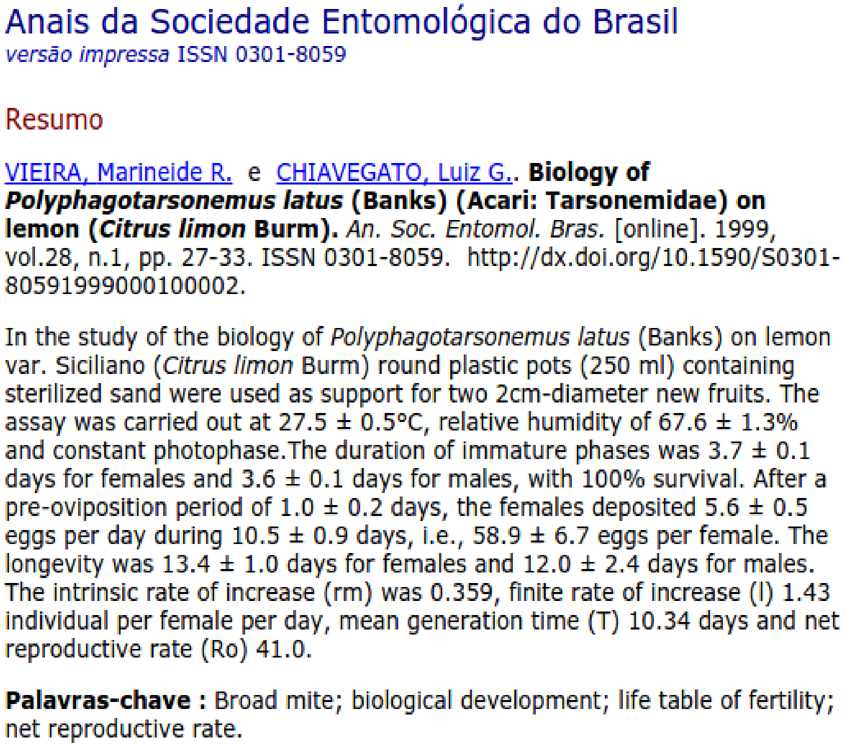
Anais da Sociedade Entomológica do Brasil
versão impressa ISSN 0301-8059

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 ± 0,5°C, umidade relativa de 67,6 ± 1,3% e fotofase contínua. O período de ovo a adulto durou 3,7 ± 0,1 dias para fêmeas e 3,6 ± 0,1 dias para machos, com sobrevivência de 100%. Após um período de pré-oviposição de 1,0 ± 0,2 dias, as fêmeas depositaram 5,6 ± 0,5 ovos por dia durante 10,5 ± 0,9 dias, totalizando 58,9 ± 6,7 ovos por fêmea. A longevidade foi de 13,4 ± 1,0 dias para fêmeas e 12,0 ± 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

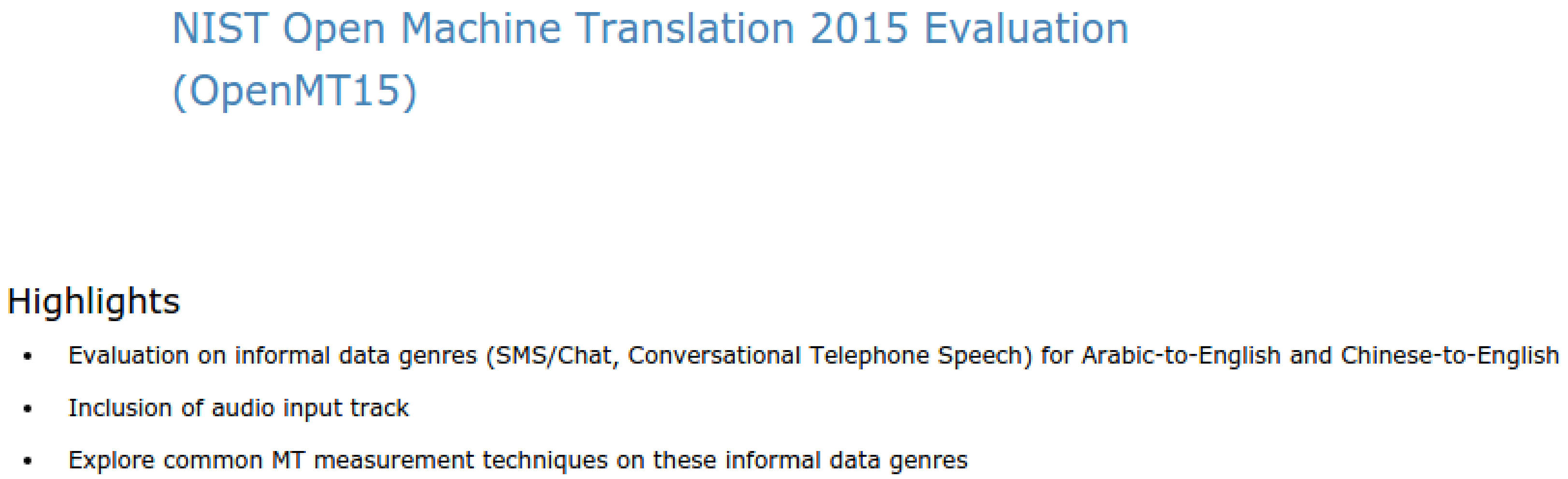
VIEIRA, Marineide R. e CHIAVEGATO, Luiz G.. **Biology of *Polyphagotarsonemus latus* (Banks) (Acari: Tarsonemidae) on lemon (*Citrus limon* Burm).** *An. Soc. Entomol. Bras.* [online]. 1999, vol.28, n.1, pp. 27-33. ISSN 0301-8059. http://dx.doi.org/10.1590/S0301-80591999000100002.

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 ± 0.5°C, relative humidity of 67.6 ± 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.

(http://www.scielo.br/)

Evaluation campaigns



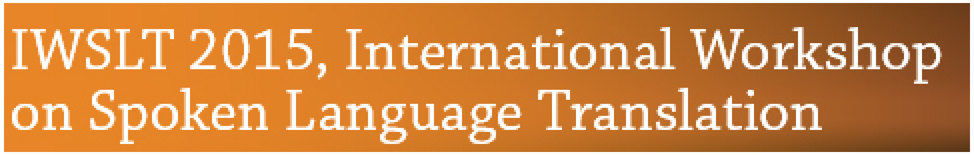
NIST Open Machine Translation 2015 Evaluation (OpenMT15)

Highlights

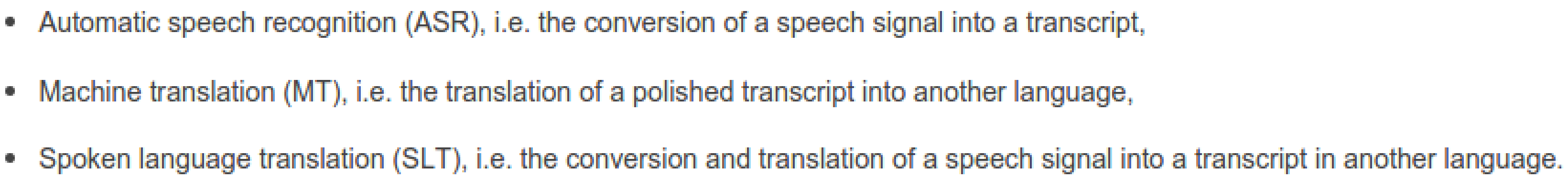
- Evaluation on informal data genres (SMS/Chat, Conversational Telephone Speech) for Arabic-to-English and Chinese-to-English
- Inclusion of audio input track
- Explore common MT measurement techniques on these informal data genres

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

Evaluation campaigns



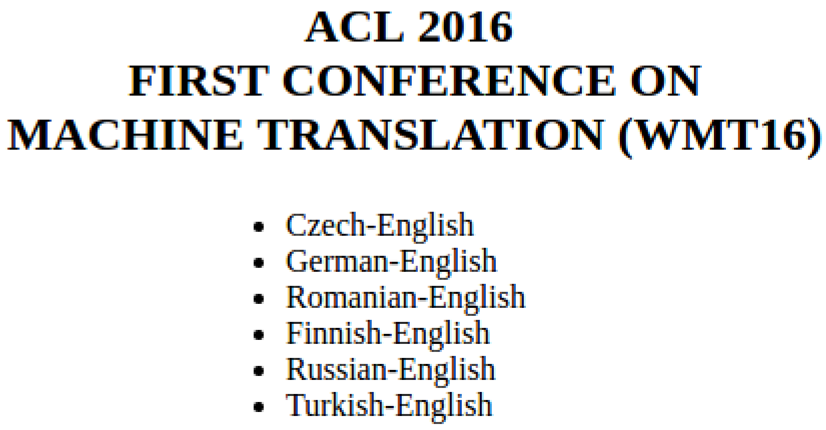
IWSLT 2015, International Workshop on Spoken Language Translation



- 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.

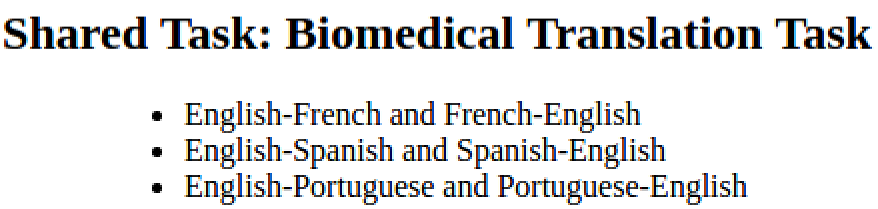
(http://workshop2015.iwslt.org/59.php)

Evaluation campaigns



ACL 2016
FIRST CONFERENCE ON
MACHINE TRANSLATION (WMT16)

- Czech-English
- German-English
- Romanian-English
- Finnish-English
- Russian-English
- Turkish-English



Shared Task: Biomedical Translation Task

- English-French and French-English
- English-Spanish and Spanish-English
- English-Portuguese and Portuguese-English

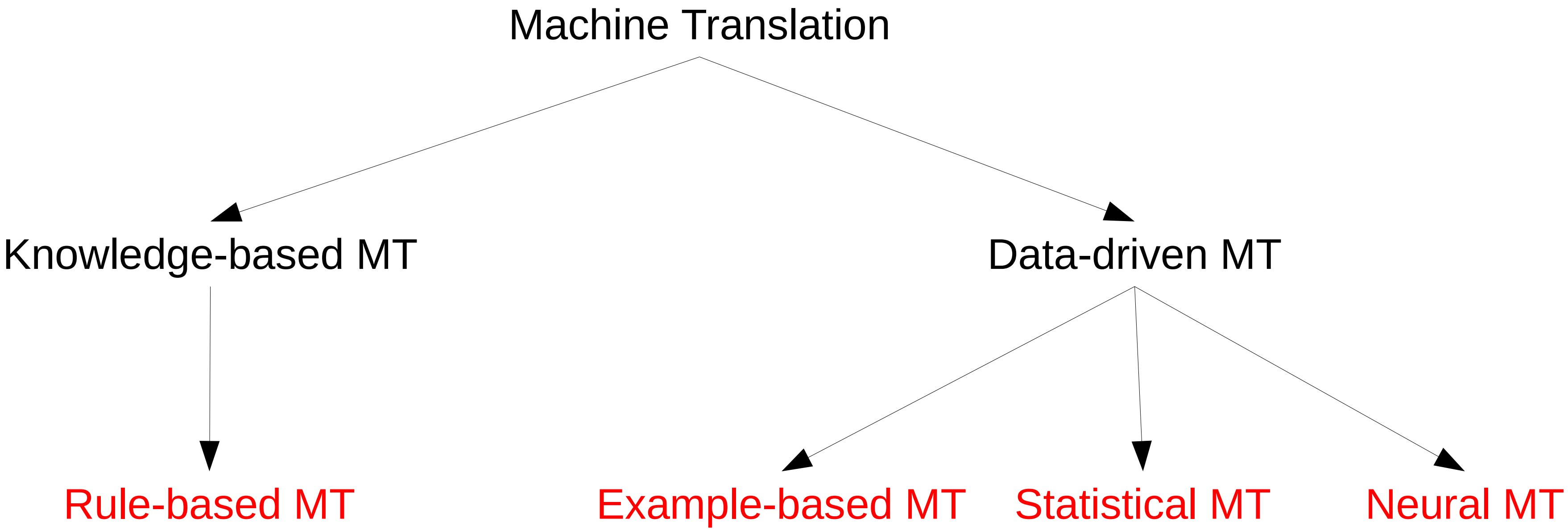
HPI at the WMT’16, WMT'17!

(http://www.statmt.org/wmt16/)

Overview

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

MT paradigms



Machine Translation

Knowledge-based MT

Data-driven MT

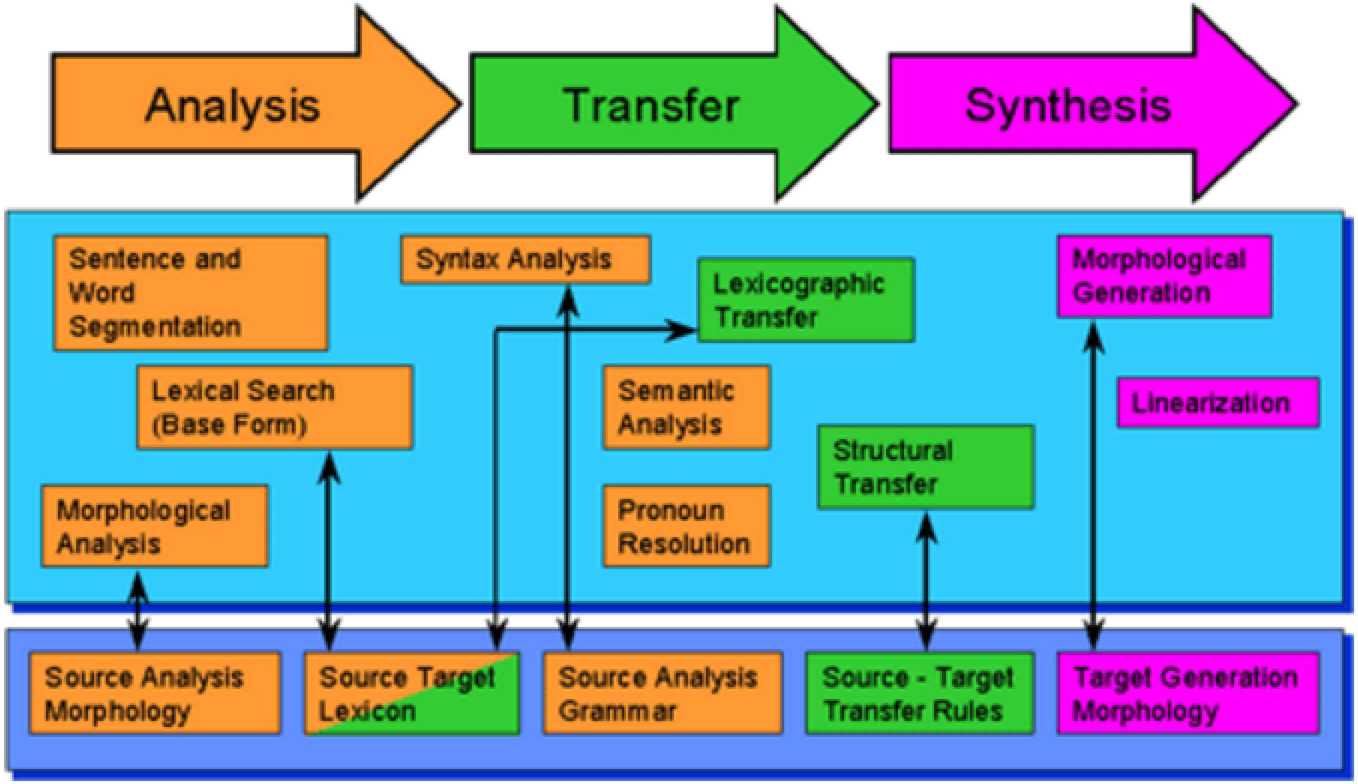
Rule-based MT

Example-based MT

Statistical MT

Neural MT

Rule-based MT

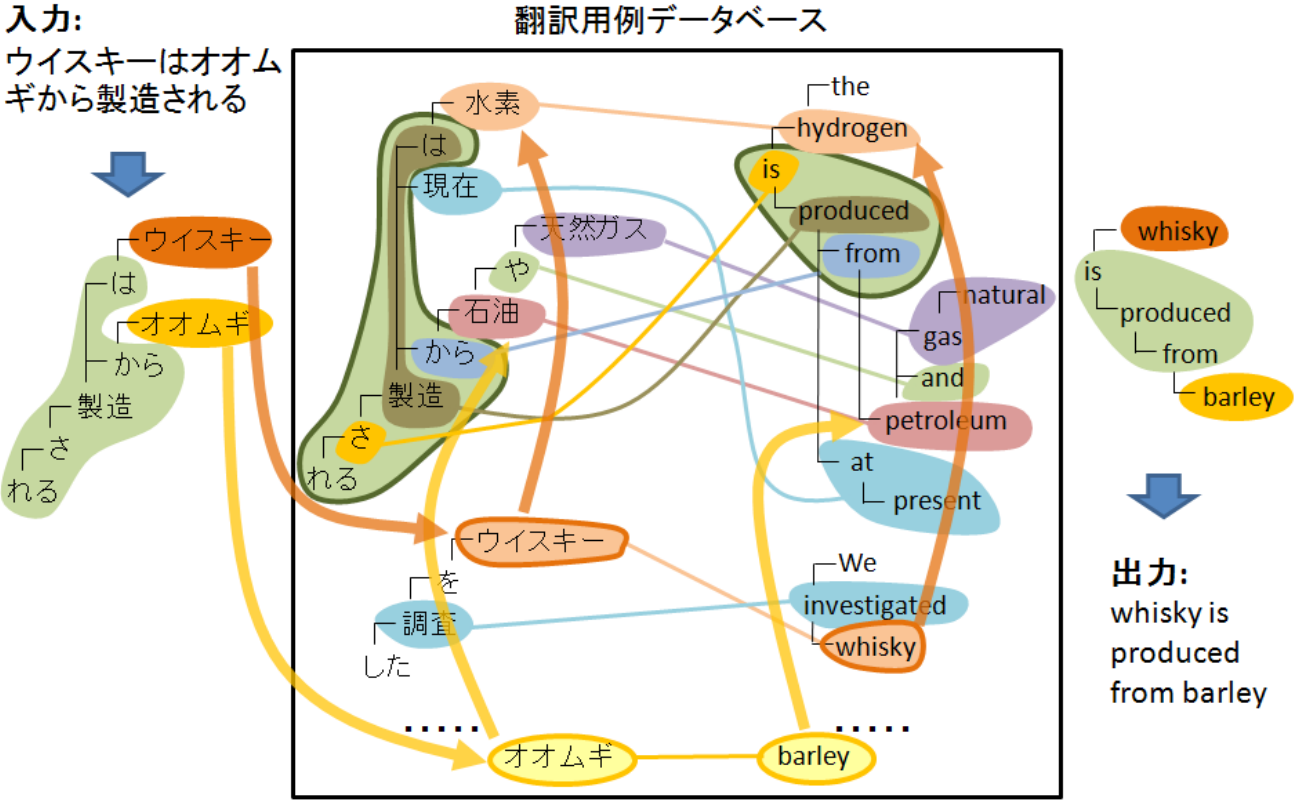


(https://nlp.fi.muni.cz/web3/en/MachineTranslation)

Rule-based MT

- Apertium (https://www.apertium.org)

Example-based MT



入力:
ウイスキーはオオムギから製造される

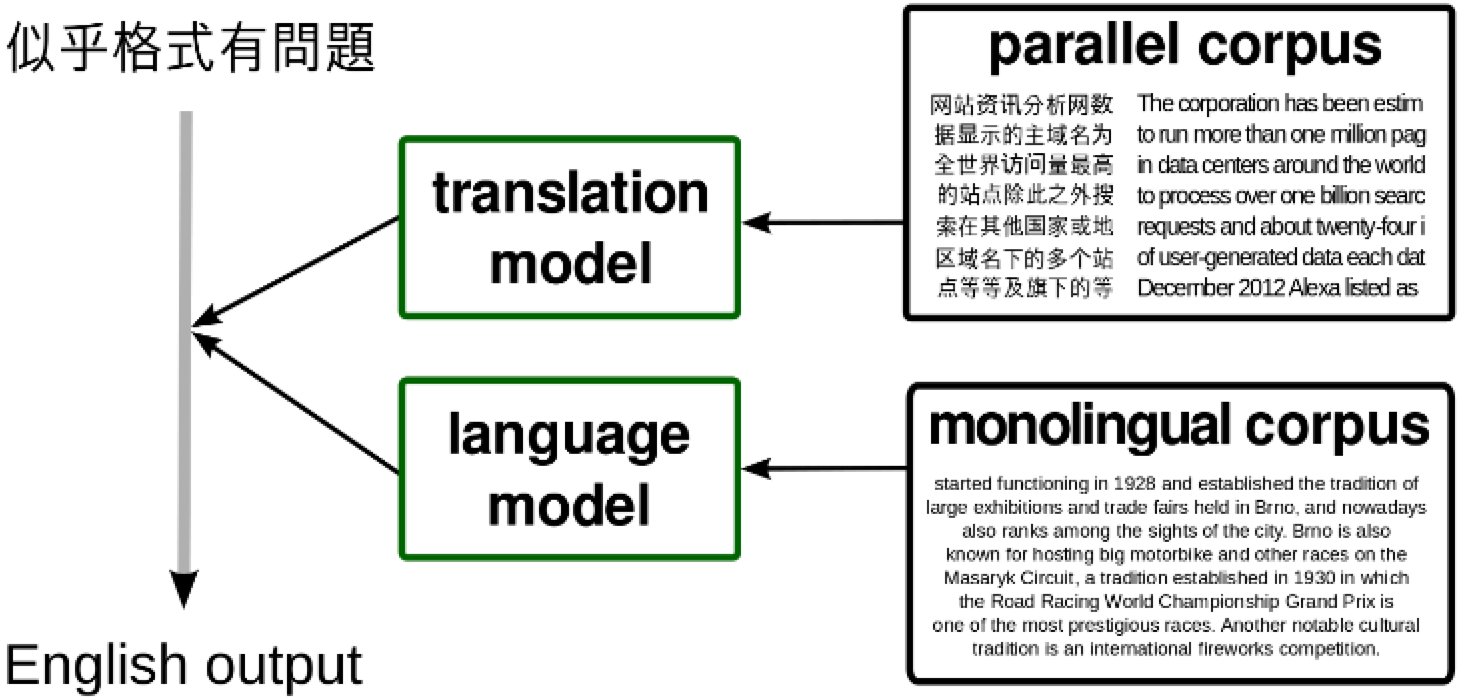
出力:
whisky is produced from barley

(http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?plugin=attach&refer=KUROHASHI-KAWAHARA-LAB&openfile=EBMT.png)

Example-based MT

- Cunei (http://cunei.sourceforge.net/)
- KyotoEBMT (http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KyotoEBMT)

Statistical MT



似乎格式有問題

translation model

language model

English output

parallel corpus

网站资讯分析网数
据显示的主域名为
全世界访问量最高
的站点除此之外搜
索在其他国家或地
区域名下的多个站
点等等及旗下的等

The corporation has been estim
to run more than one million pag
in data centers around the world
to process over one billion searc
requests and about twenty-four i
of user-generated data each dat
December 2012 Alexa listed as

monolingual corpus

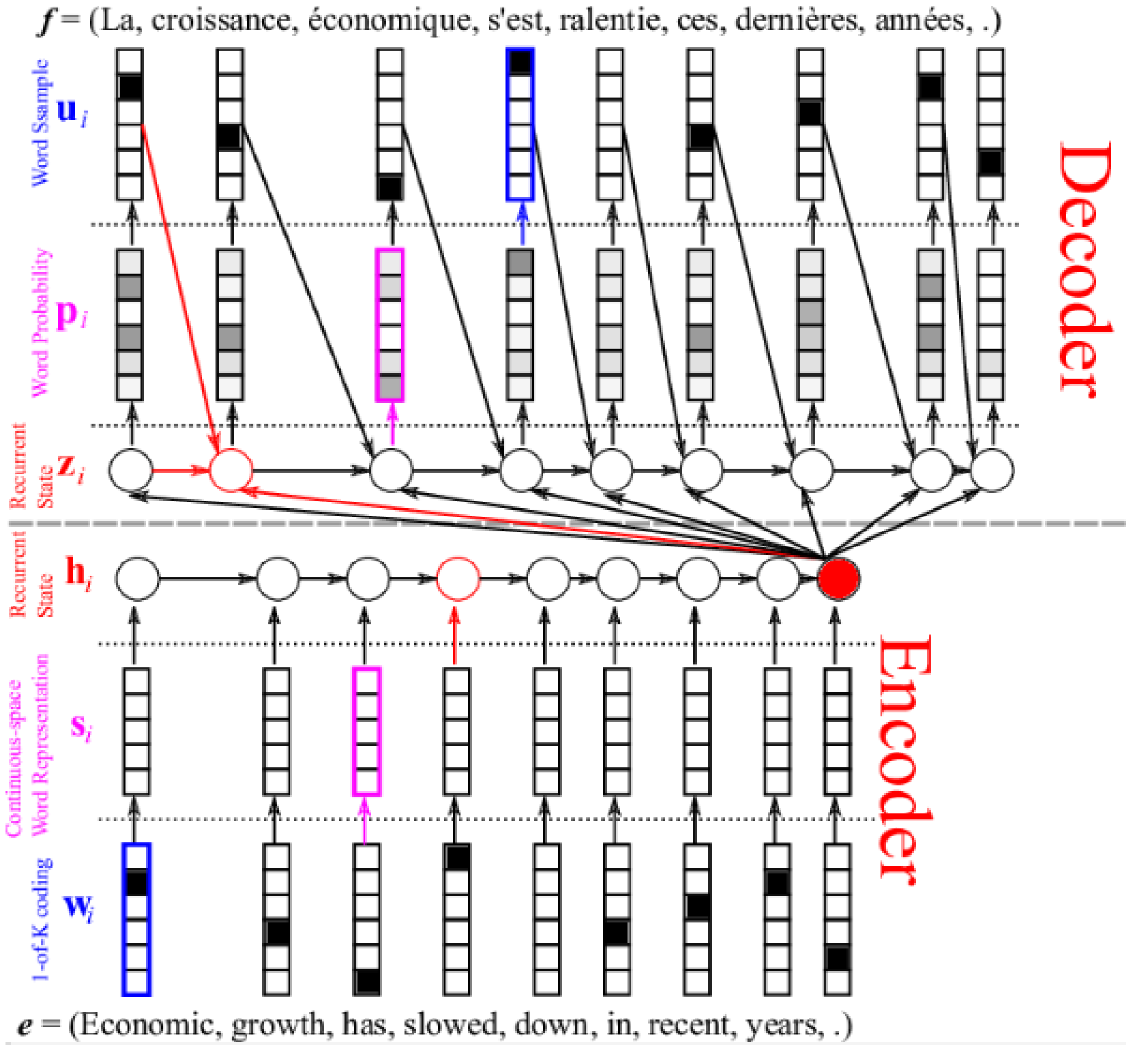
started functioning in 1928 and established the tradition of large exhibitions and trade fairs held in Brno, and nowadays also ranks among the sights of the city. Brno is also known for hosting big motorbike and other races on the Masaryk Circuit, a tradition established in 1930 in which the Road Racing World Championship Grand Prix is one of the most prestigious races. Another notable cultural tradition is an international fireworks competition.

(https://nlp.fi.muni.cz/web3/en/MachineTranslation)

Statistical MT

- Moses (http://www.statmt.org/moses/)
- Cunei (http://cunei.sourceforge.net/)

Neural MT



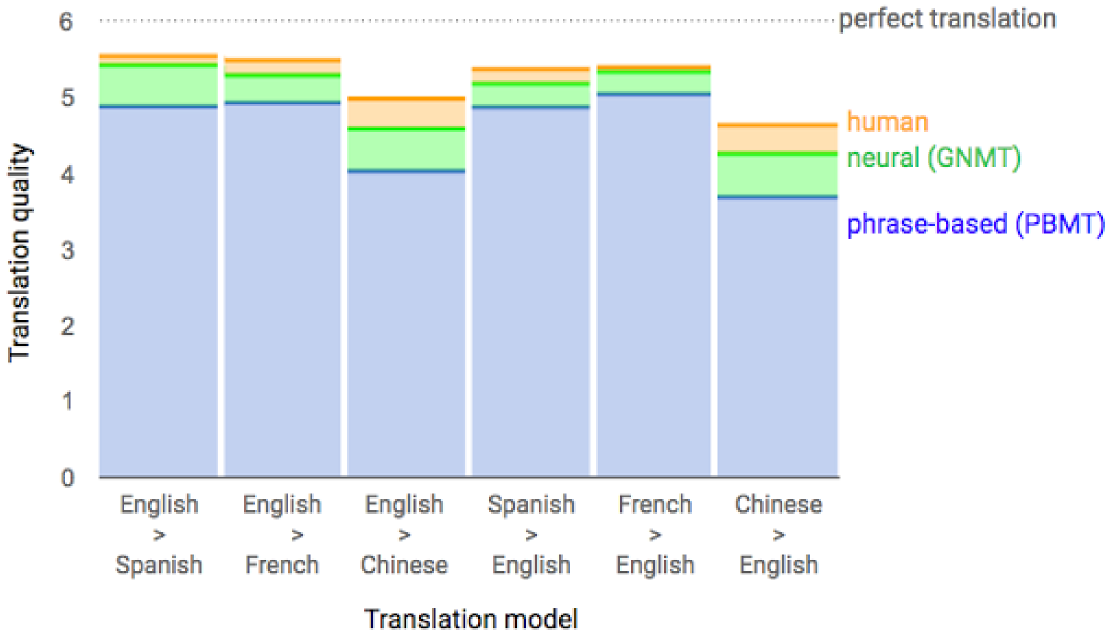
f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
$\mathbf{u}_i$
Word Probability
$\mathbf{p}_i$
Recurrent State
$\mathbf{z}_i$
Decoder
Recurrent State
$\mathbf{h}_i$
Continuous-space Word Representation
$\mathbf{s}_i$
1-of-K coding
$\mathbf{w}_i$
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Figure 2. The very first neural machine translation system.

(https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/)

Neural MT

- LISA (http://104.131.78.120/)
- TensorFlow (https://research.googleblog.com/2016/09/a-neural-network-for-machine.html)



Overview

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

MT course – what to expect from me

- Overview on MT methods

- Supervision of the projects

- Be available by email and in the office (Villa room 0.01)

MT course – what I expect from you

- Presence and participation in the lecture (not controlled)
- Take part in a project (team or individual)
- Take part in the final exam

Project

- Teams of 2/3 students
- „Take part“ in one of the translation challenges at WMT‘16 (http://www.statmt.org/wmt16/)
  - News
  - IT-domain
  - Biomedical
- Presentation of preliminary and final results
- Submission of a 3-pages report
- Source code in GitHub or similar

Project

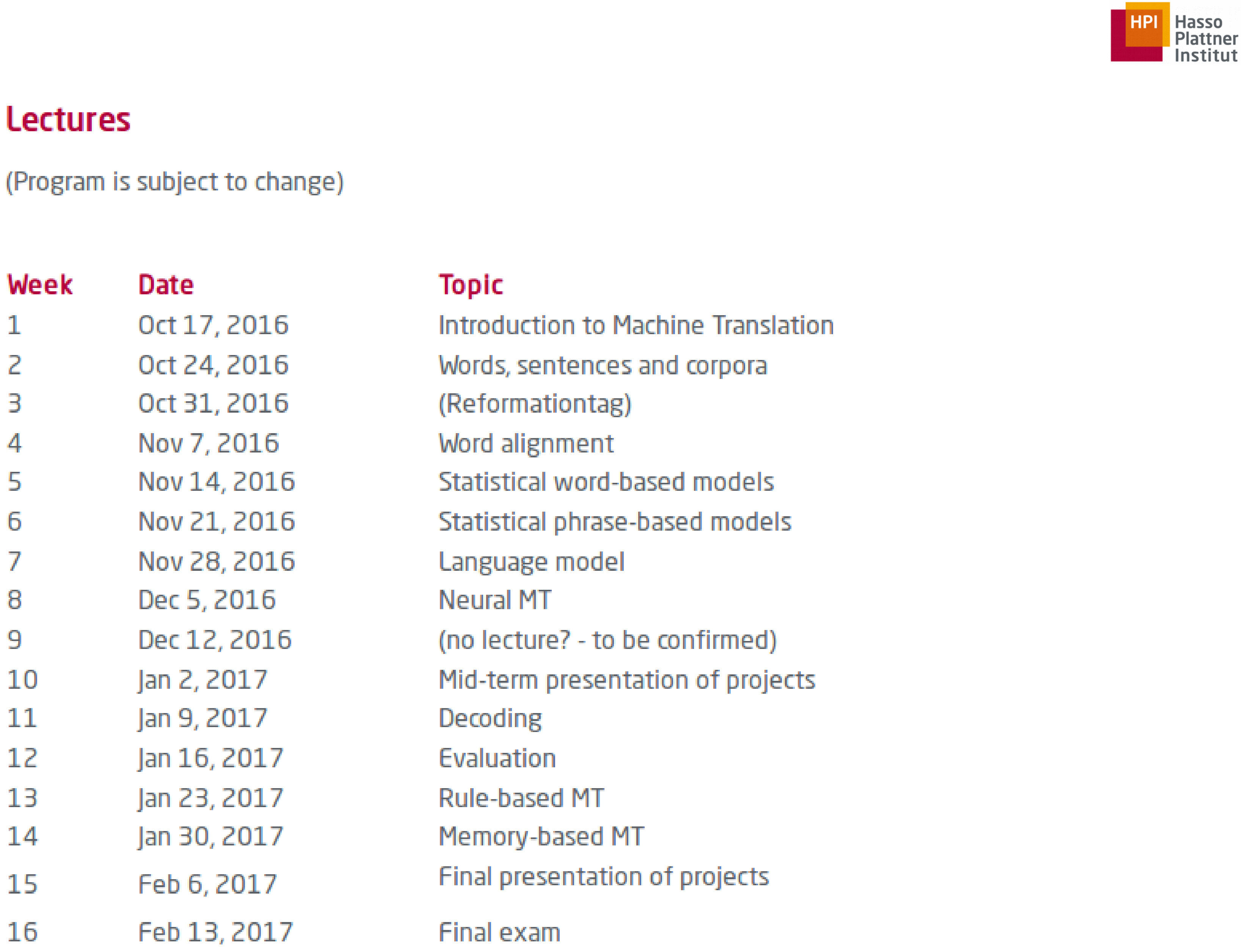
- Flexible...
  - „Any“ translation task (first-come, first-served)
  - Any language pair
  - Any MT paradigm
  - Any NLP/MT tools

Project

- ...but with some requirements
  - Integration of domain-specific resources
  - Training on out-of-domain corpora (talk to other teams)
  - Evaluation of official test datasets (last year‘s test data)

Project

- Mail to me (mariana.neves@hpi.de):
  - Team members
  - WMT translation task(s)
  - Language pair(s)
  - Host of the project (GitHub, etc)



Lectures

(Program is subject to change)

Week	Date	Topic
1	Oct 17, 2016	Introduction to Machine Translation
2	Oct 24, 2016	Words, sentences and corpora
3	Oct 31, 2016	(Reformationtag)
4	Nov 7, 2016	Word alignment
5	Nov 14, 2016	Statistical word-based models
6	Nov 21, 2016	Statistical phrase-based models
7	Nov 28, 2016	Language model
8	Dec 5, 2016	Neural MT
9	Dec 12, 2016	(no lecture? - to be confirmed)
10	Jan 2, 2017	Mid-term presentation of projects
11	Jan 9, 2017	Decoding
12	Jan 16, 2017	Evaluation
13	Jan 23, 2017	Rule-based MT
14	Jan 30, 2017	Memory-based MT
15	Feb 6, 2017	Final presentation of projects
16	Feb 13, 2017	Final exam

(https://hpi.de//en/plattner/teaching/winter-term-201617/machine-translation.html)

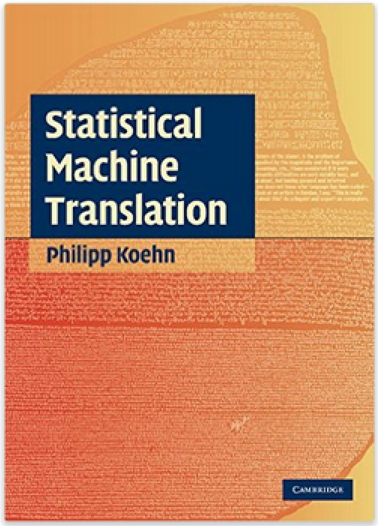
Grading

- 60% Project
  - Commitment, implementation, presentation, report
  - Each team member should present in either of the two appointments (mid-term or final)

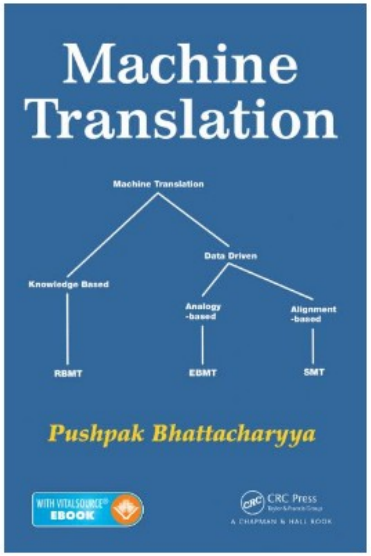
- 40% Final exam

Course books

- Statistical Machine Translation
  - Philipp Koehn

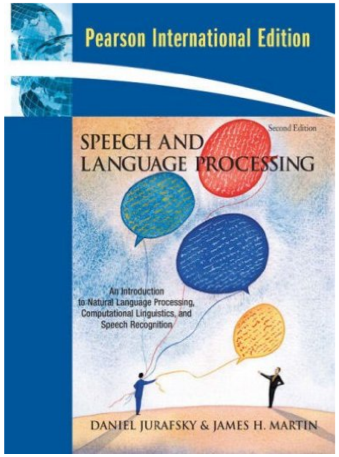


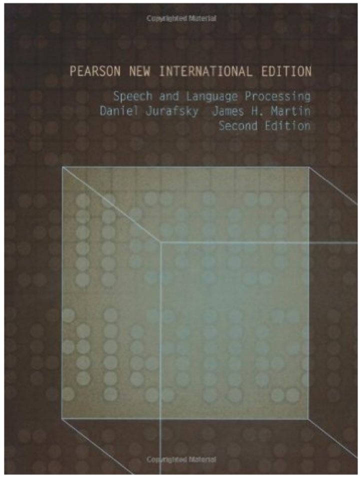
- Machine Translation
  - Pushpak Bhattacharyya



Course books

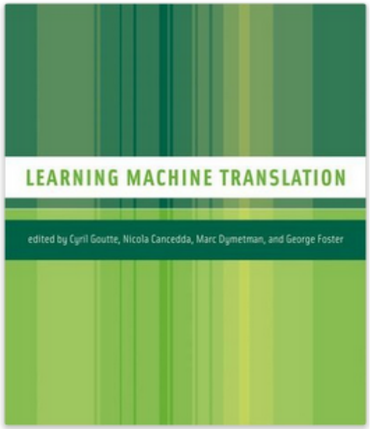
- Speech and Language Processing (Chapter 25)
  - Daniel Jurafsky and James H. Martin


Pearson International Edition
Second Edition
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN

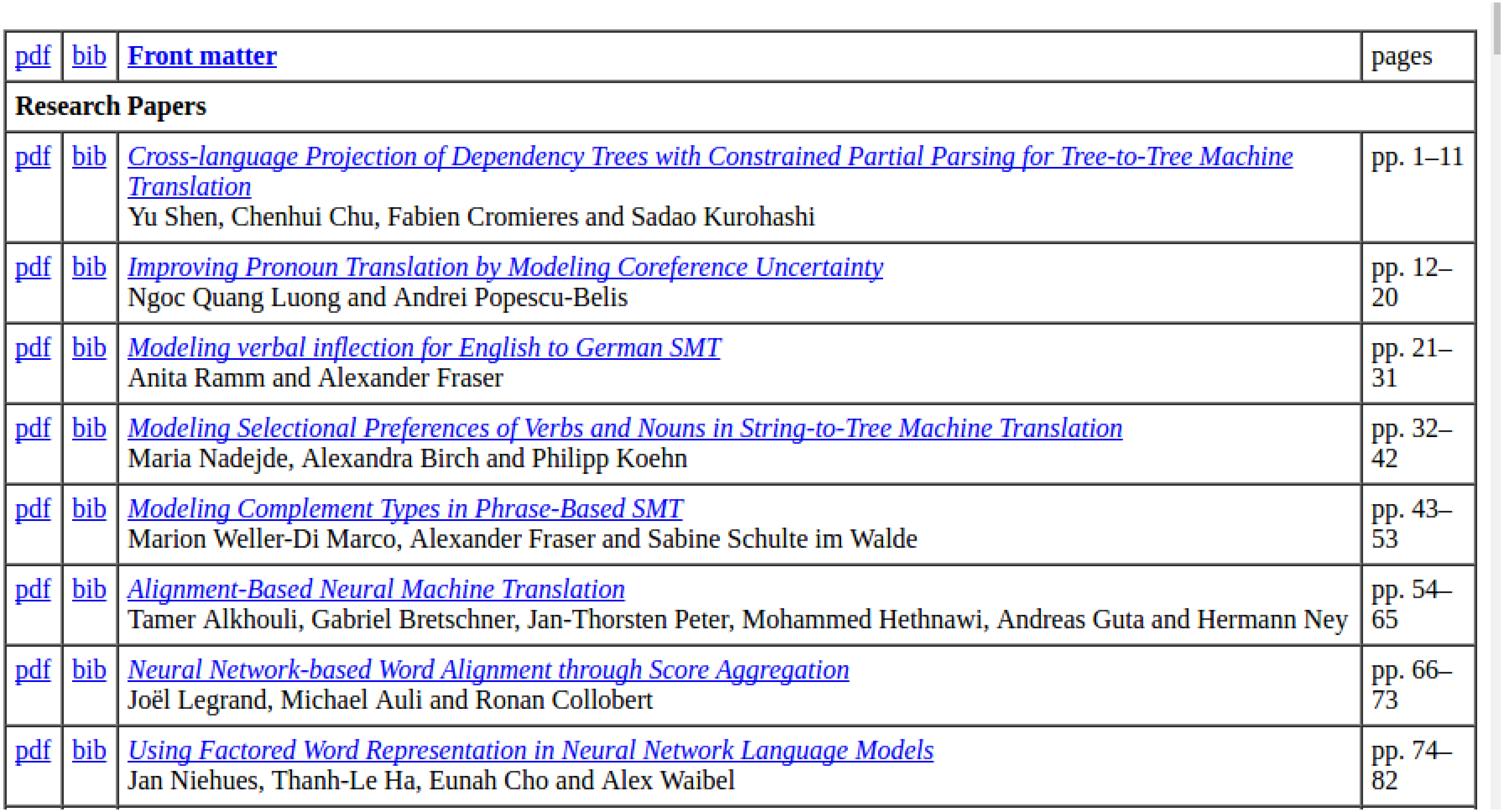

Copyrighted Material
PEARSON NEW INTERNATIONAL EDITION
Speech and Language Processing
Daniel Jurafsky James H. Martin
Second Edition
Copyrighted Material

Course books (advanced topics)

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


LEARNING MACHINE TRANSLATION
edited by Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster

Workshop papers



pdf	bib	Front matter	pages
Research Papers			
pdf	bib	*Cross-language Projection of Dependency Trees with Constrained Partial Parsing for Tree-to-Tree Machine Translation* Yu Shen, Chenhui Chu, Fabien Cromieres and Sadao Kurohashi	pp. 1–11
pdf	bib	*Improving Pronoun Translation by Modeling Coreference Uncertainty* Ngoc Quang Luong and Andrei Popescu-Belis	pp. 12–20
pdf	bib	*Modeling verbal inflection for English to German SMT* Anita Ramm and Alexander Fraser	pp. 21–31
pdf	bib	*Modeling Selectional Preferences of Verbs and Nouns in String-to-Tree Machine Translation* Maria Nadejde, Alexandra Birch and Philipp Koehn	pp. 32–42
pdf	bib	*Modeling Complement Types in Phrase-Based SMT* Marion Weller-Di Marco, Alexander Fraser and Sabine Schulte im Walde	pp. 43–53
pdf	bib	*Alignment-Based Neural Machine Translation* Tamer Alkhouli, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta and Hermann Ney	pp. 54–65
pdf	bib	*Neural Network-based Word Alignment through Score Aggregation* Joël Legrand, Michael Auli and Ronan Collobert	pp. 66–73
pdf	bib	*Using Factored Word Representation in Neural Network Language Models* Jan Niehues, Thanh-Le Ha, Eunah Cho and Alex Waibel	pp. 74–82

(http://www.statmt.org/wmt16/papers.html)