Machine Translation
Computer–Aided Translation

Machine Language Processing

Martin Kappus (kapm@zhaw.ch)

Machine Translation
Computer-Aided Translation

Agenda

Machine Translation

- Introduction
- History
- Approaches

Computer-Aided Translation

- Introduction
- Hands-on Project

Human and machine translation

Machine Translation

- MT as a model of human translation
- MT as automation a of human activity
  - Fully automated high quality translation
  - Human-aided machine translation
  - Machine-aided human translation
  - Conventional human translation

Human and machine translation

human involvement →

← *mechanization*

fully automatic high-quality translation (FAHQT)

human-aided machine translation (HAMT)

machine-aided human translation (MAHT)

traditional human translation

Computer-Assisted Translation (CAT)

Histroy of the development of MT

Ancestors

- Leibniz, Descartes (lexical equivalents of all known languages would be given the same code number)
- Joh. J. Becher 1661
- John Wilkins 1668 (*Essay towards a Real Character and a Philosophical Language*)

Precursors

- George Artstruni 1933
- Petr P. Smirnov-Trojanskij 1933

Pioneers

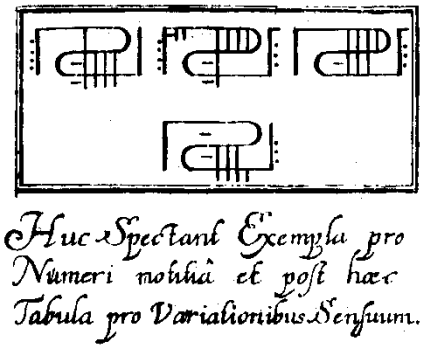
- Warren Weaver, mathematician, *Translation* (1949)
- Yehoshua Bar-Hillel, logician

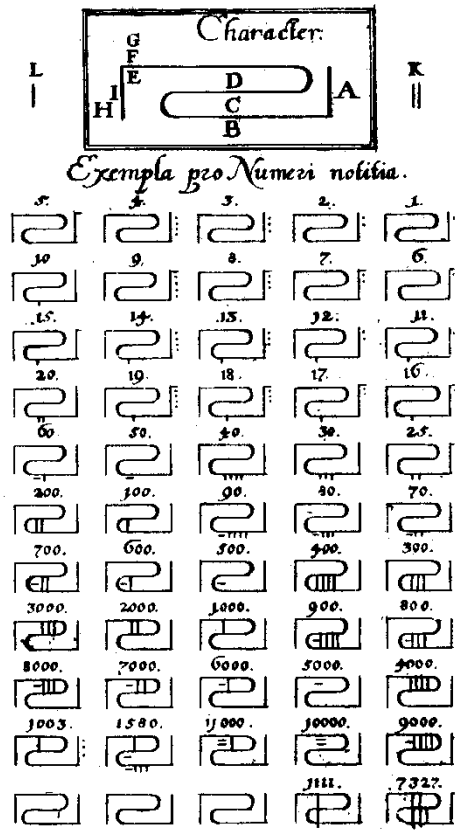
Ancestors: Johannes Becher

JOH. J. BECHERI,
Spirensis
CHARACTER,
Pro
NOTITIA LIN-
guarum Universali.
INVENTUM STEGANO-
GRAPHICUM HACTENUS IN-
auditum quo quilibet suam Legendo vernaculam
diversas imò omnes Linguas, unius etiam diei
informatione, explicare ac intelli-
gere potest.

FRANCOFURTI,
Sumpt. Johannis Wilh. Ammonii
& Wilhelmi Serlini,
Typis JOHANNIS GEORGII SPÖRLIN.
ANNO M. DC. LXI.

Ancestors: Johannes Becher





Encoding

Ancestors: Johannes Becher

LEXICON
Pro
Resolutione primæ Characteris
partis
A. B. C. D.
L. S.

Ineas Characteris secundùm informationem in tuæ Vernaculæ numerum transfer quem in ejusdem LEXICO evolve, ita designatum vocabulum pro conditione punctationis flectendum invenies.

A Ab	1	Abbatissa	10
Abactus	3	Abdere	11
Abacus	4	Abdicare	12
Abaculus	5	Abditus	13
Abavus	6	Abdomen	14
Abax	7	Aberrare	15
Abbas	8	Abesse	16
Abbatia	9	Abhinc	17
		Abhorrere	18

Abies

CHARACTERIS.

Accipiter	83	Acidus	115
Acclivis	84	Acies	116
Accola	85	Acinaces	117
Accommodare	86	Acinus	118
Accumbere	87	Acinosus	119
Accumlare	88	Acipenser	120
Accurate	89	Aclis	121
Accusare	90	Aconitum	122
Accusavi	91	Acorus	123
Accusabo	92	Acquiescere	124
Accusa	93	Acquirere	125
Accusem	94	Acredula	126
Accusaverim	95	Acris, cre	127
Accusavero	96	Acriter	128
Accusavisse	97	Acta, orum	129
Accusatum ire	98	Actio	130
Accusans	99	Actor	131
Accusatio	100	Actum est	132
Acedia	101	Actuarius	133
Acer	102	Actuosus	134
Acer, acris	103	Actus	135
Acerbus	104	Actutum	136
Aceria	105	Acuere	137
Acervus	106	Aculeus	138
Acervare	107	Acumen	139
Acetabulum	108	Acupictus	140
Acetarium	109	Acus, eris	141
Acetosa	110	Acus	142
Acetosella	111	Acufacere	143
Acetum	112	Acupingere	144
Achates	113	Acutus	145
Acicula	114	Ad	146

C Ad

Lexicon

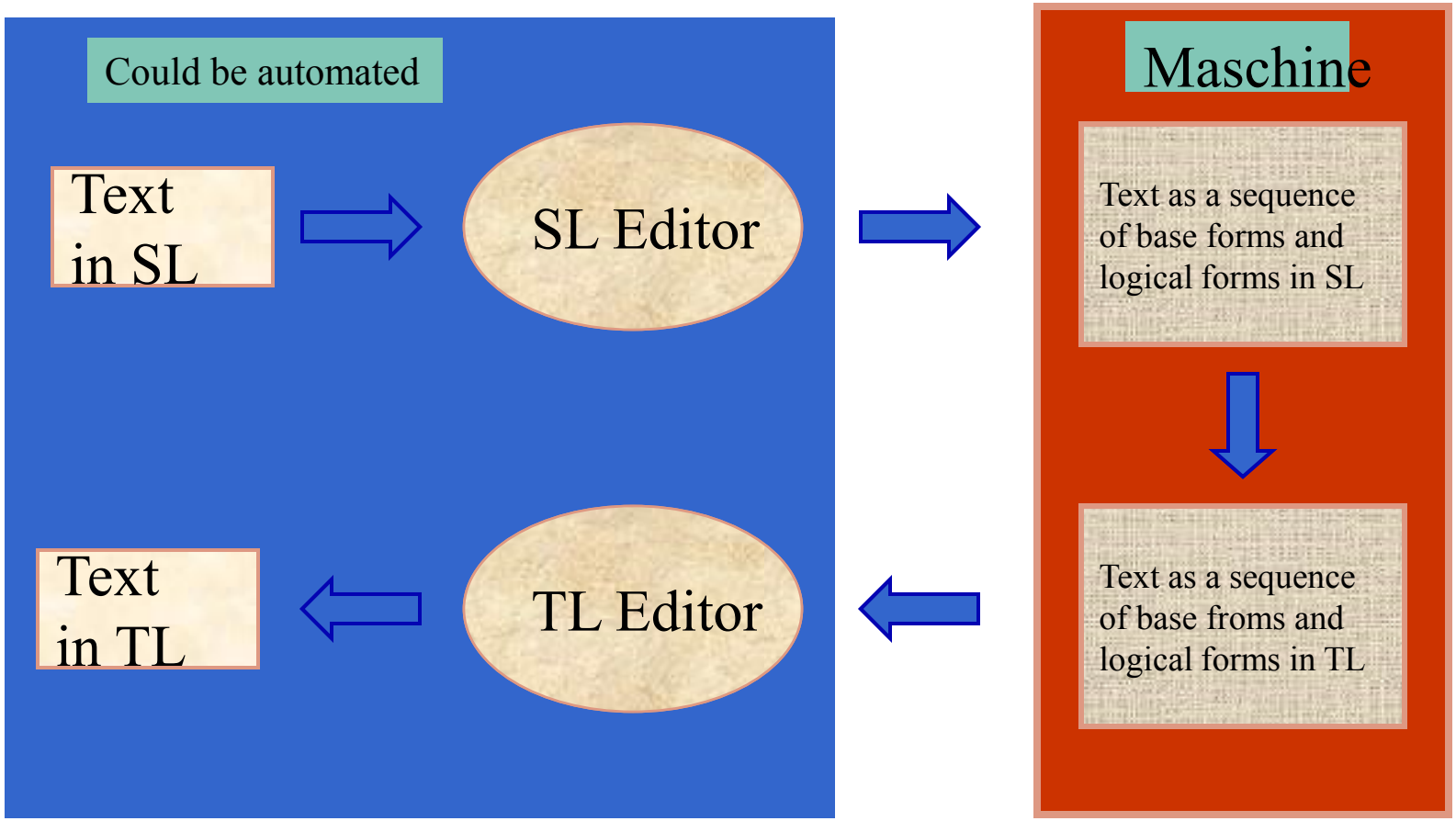
Precursors: Georges Artsrouni (1933)

- Georges Artsrouni developed a translation machine which he called a “Mechanical Brain”.
- It consisted of a storage device (paper) that could find the translation for any given word in another language.
- A patent was issued in 1933 and a prototype machine was exhibited and demonstrated in 1937.

Precursors: Petr Smirnov-Troyanskii

- Patent issued in September 1933 to Smirnov-Troyanski: "construction of a “machine for the selection and printing of words while translating from one language into another or into several others simultaneously.”
- Troyanskii envisaged three stages in the translation process;
  1. A human editor knowing only the source language was to analyze the input text into a particular ‘logical’ form.
  2. In the second stage the machine was designed to transform sequences of base forms and ‘logical symbols’ of source texts into sequences of base forms and symbols of target languages.
  3. In the third stage an editor knowing only the target language was to convert this sequence into the normal forms of his own language.
- Troyanskii believed that the process of logical analysis could itself be mechanized, by means of a machine specially constructed for the purpose”

Precursors: Petr Smirnov-Troyanskij



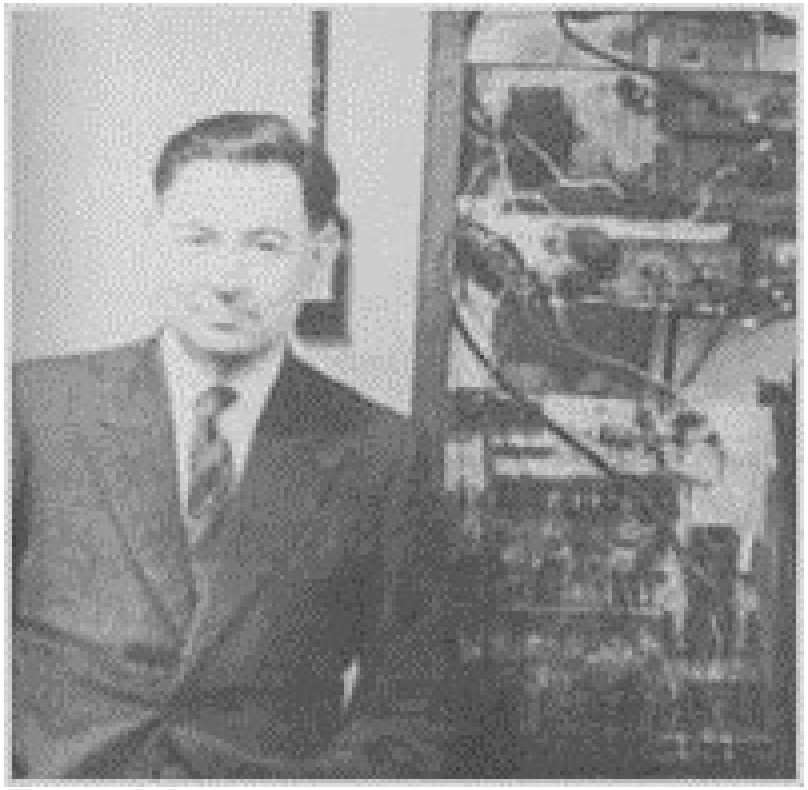
Pioneers

- Within a few years of the first appearance of the ‘electronic calculators’ research had begun on using computers as aids for translating natural languages.
- Attributed to conversations between Warren Weaver (Rockefeller Foundation) and Andrew D. Booth
- Within a few years research on machine translation (MT) had begun at many US universities (Washington U, UCLA, MIT
- In 1951 Yeoshua Bar-Hillel was appointed to first researcher solely dedicated to MT

Pioneers

- In 1954 the first public demonstration of the feasibility of machine translation was given (a collaboration by IBM and Georgetown University).
  (see http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf)
- Although using a very restricted vocabulary and grammar it was sufficiently impressive to stimulate massive funding of MT in the United States and to inspire the establishment of MT projects throughout the world

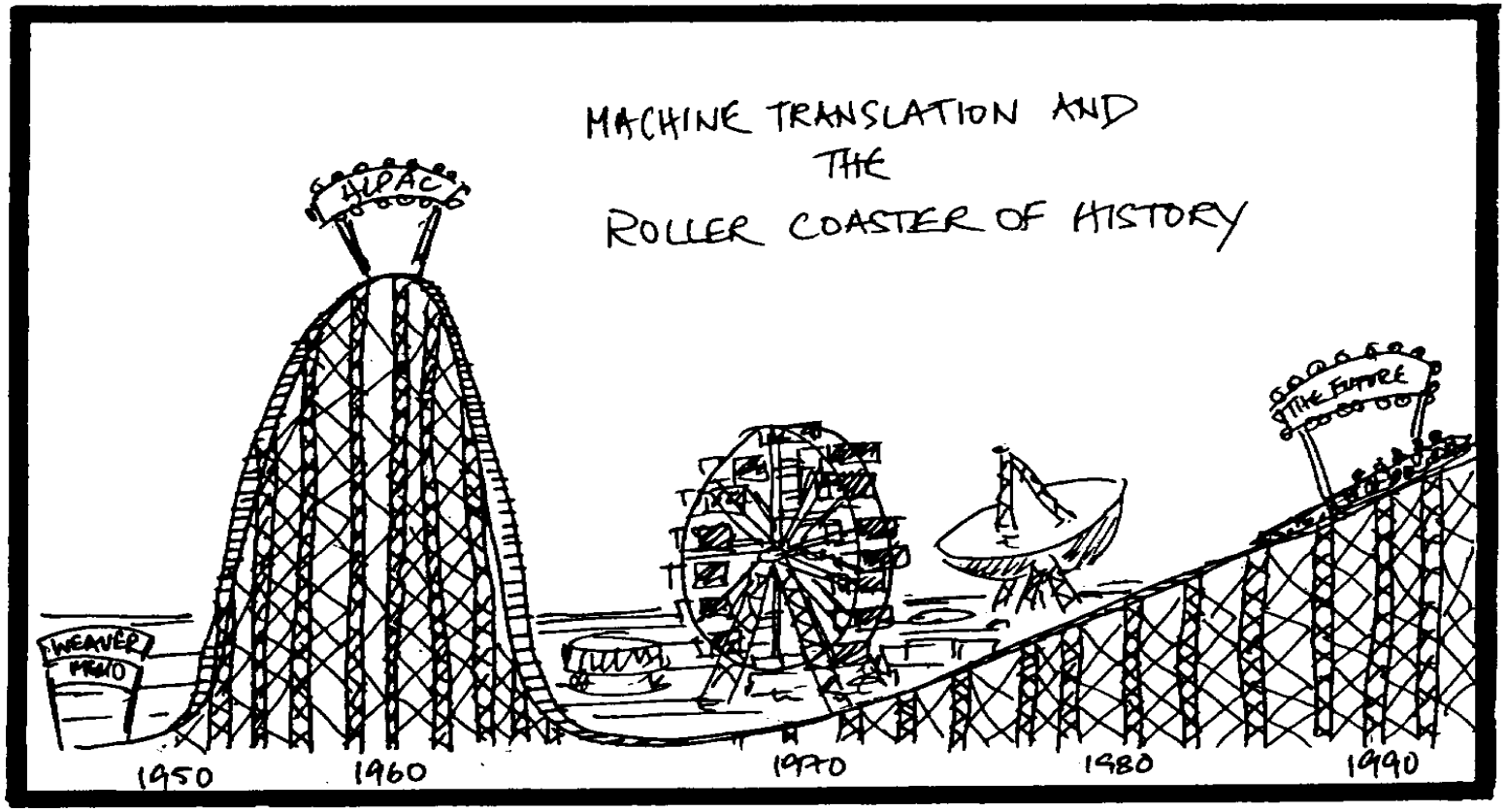
Pioneers : Warren Weaver and Andrew Booth



Weaver Memorandum

- The idea of MT was promoted by a memorandum by Warren weaver in 1949 (Weaver 1949).
- Weaver's memorandum concentrated more on the general strategies and long-term objectives of MT than on the more technical problems Booth and others had been tackling.
- Weaver raised four interesting theoretical points:
  - the problem of multiple meaning,
  - the logical basis of language,
  - the application of communication theory and cryptographic techniques,
  - The possibilities of language universals.

MT and the Rolle Coaster of History



History of the development of MT

The decade of optimism: 1954-1966

- earliest systems consisted primarily of
  - large bilingual dictionaries
  - some rules for producing the correct word order in the output
- the need for more systematic methods of syntactic analysis became evident
- a number of projects were inspired by contemporary developments in linguistics (generative grammar)
- by 1964, the US government became concerned at the lack of progress → the Automatic Language Processing Advisory Committee (ALPAC),

Histroy of the development of MT

The aftermath of the ALPAC report: 1966-1980

- "there is no immediate or predictable prospect of useful machine translation"
- the ALPAC report brought a virtual end to MT research in the United States for over a decade
- research continued in Canada, in France and in Germany
- in the 70s several operational systems appeared.
- From the mid-1970s onwards the demand for MT came from administrative and commercial demands of multilingual communities and multinational trade

Histroy of the development of MT

The 1980s:

- emergence of a wide variety of MT system types
- availability of microcomputers and of text-processing software created a market for cheaper (and "smaller") MT systems
- the dominant strategy was indirect translation

Histroy of the development of MT

The early 1990s:

- experiments on a system based purely on statistical methods.
- use of methods based on corpora of translation examples
- In both approaches no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents

Histroy of the development of MT

Now?

- Question is not whether to use MT but rather how to use MT in translation as a service
- Mostly HAMT and MAHT
- Translators increasingly work as post-editors as well
- Still skepticism among translators about MT
- How can MT be integrated in the typical translation workflow?

Development stages

- Computer science
- Computational linguistics
- Artificial intelligence
- Corpus linguistics

Development stages: Computer science

- MT as a engineering problem
- narrow empirical approach
- naïve linguistic approach bundled with complex coding
- no modularity
- no separation between linguistic data and process algorithms

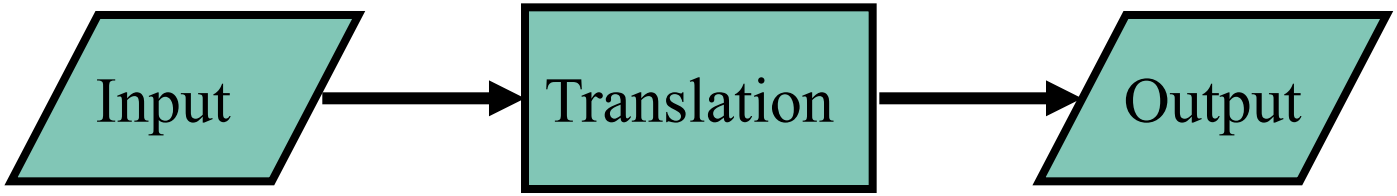
Development stages: Computational linguistics

- Increasing influence of linguistic theories
- MT as a task for the domain of computational linguistics
- Independence of analysis and synthesis (Modularity)
- Separation of linguistic data and process algorithms
- Indirect translation algorithms via stratification and transfer (possibly interlingua)

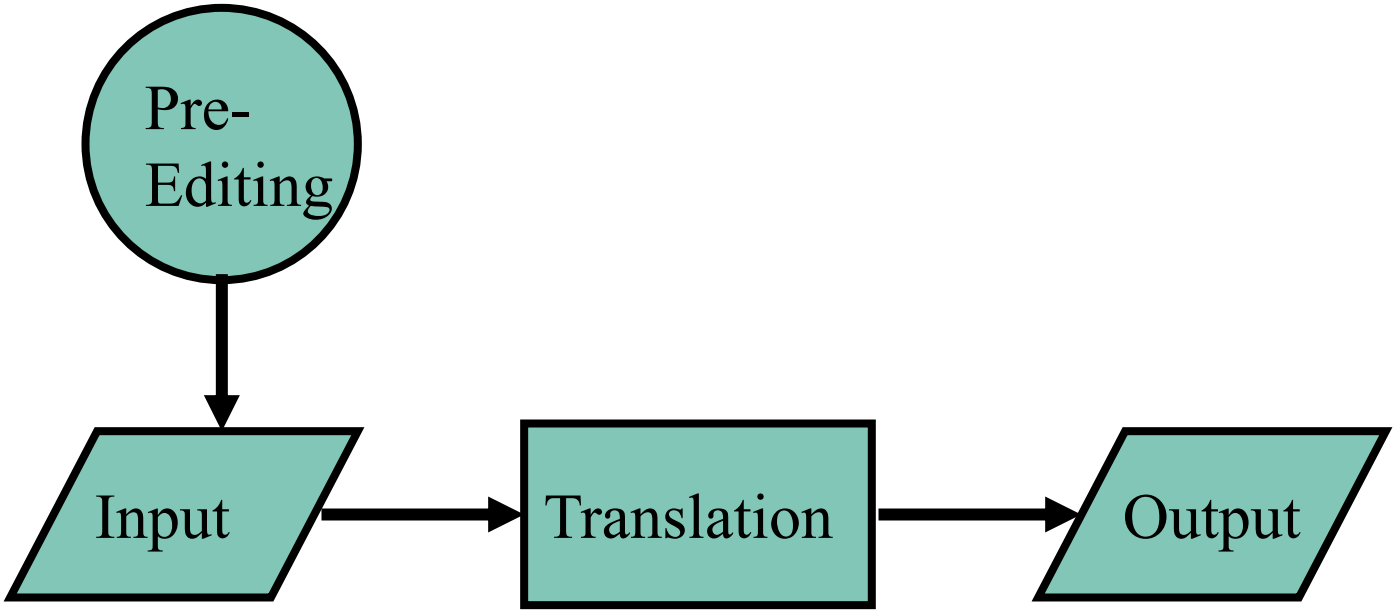
Development stages Artificial intelligence

- Inclusion of background information
- Context
- Use of cognitive schemas
- World knowledge

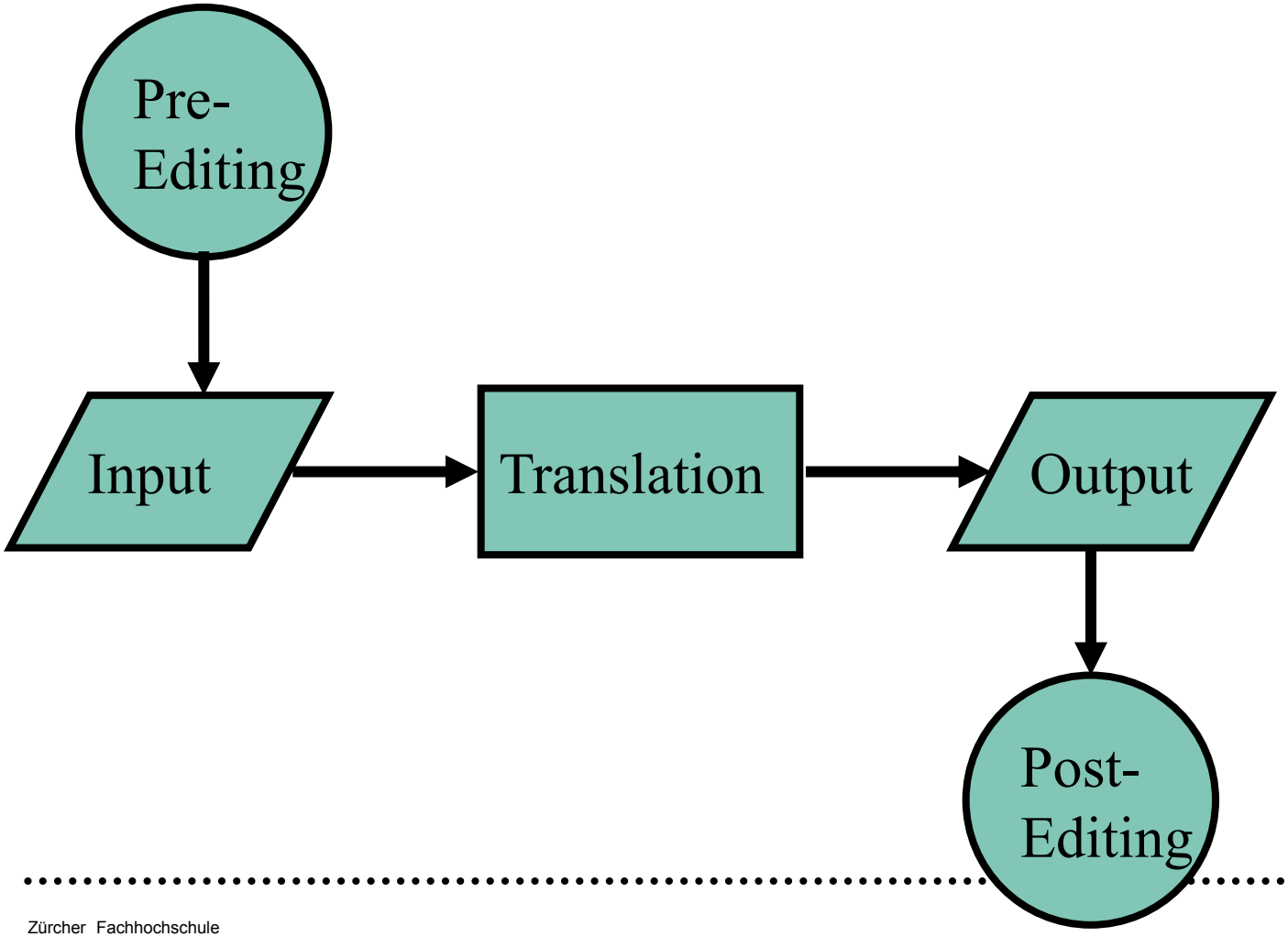
Architectures of Translation Systems: automated



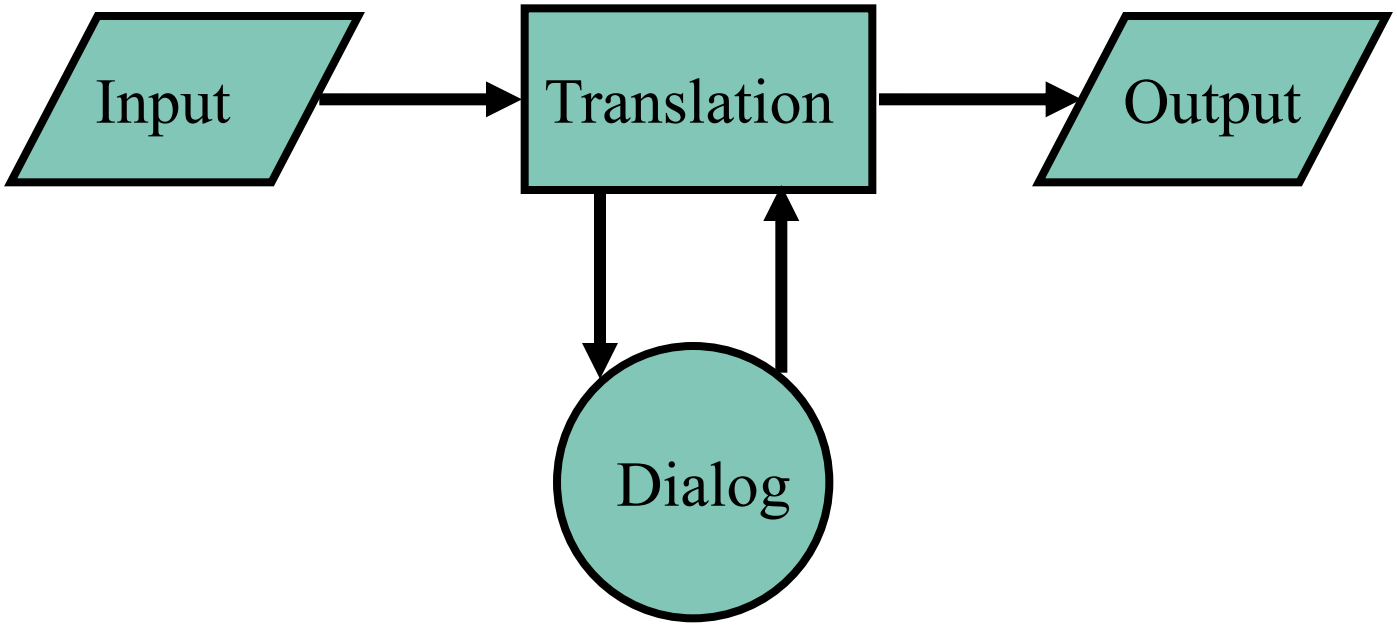
Architectures of Translation Systems: Pre-editing



Architectures of Translation Systems: Post-Editing

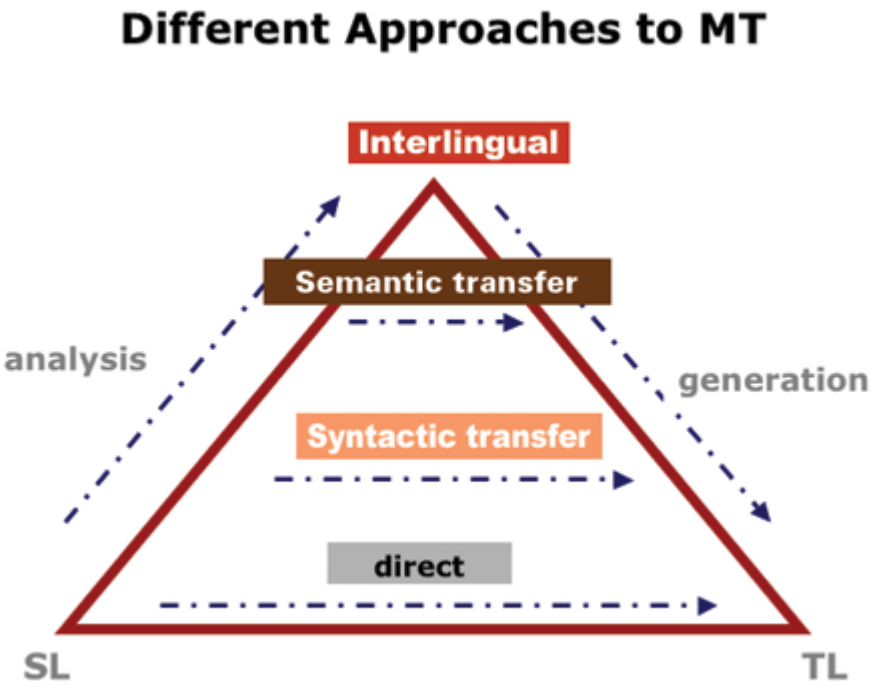


Architectures of Translation Systems: Dialog System



Direct Systems/Transfer systems/Interlingual systems

The Vauquois triangle



Source http://www.multilingual.com/articleDetail.php?id=1082

Human and machine translation

human involvement →

← *mechanization*

fully automatic high-quality translation (FAHQT)

human-aided machine trans-lation (HAMT)

machine-aided human translation (MAHT)

traditional human translation

Computer-Assisted Translation (CAT)

Computer- Aided Translation

Initial questions:

- What types of text make up the majority of all (paid) translations?
- What do you know about CAT-Tools?

Computer- Aided Translation

CAT Tools: 2 Definitions

"Generic term for systems and technologies that support the translator during the translation process."
Massion (2005)

"Software tools that allow the translator to create, to use and to maintain multilingual lexicon-databases, and text-databases.
Volk and Jekat (2010)

Computer- Aided Translation

CAT ≠ MT

CAT is not the same as machine translation (MT):

- **MT performs the translation task for the translators**
- **CAT Tools support the translators in performing their tasks**

Computer- Aided Translation

Target Audience - Who is/should be using CAT Tools?

- Professional Translators
- Translation departments in companies (manufacturing, banking, finance, administration, …)
- Language Service Providers (Translation agencies)
- Freelance Translators

Computer- Aided Translation

How can CAT Tools help to increase efficiency and to reduce cost thus allowing for a higher turn-around?

CAT Tools contain a translation memory (TM)

Translations are saved in the TM together with the source text.

When the a sentence occurs that has already been translated (or a similar sentence) the stored thranslation are suggested to the translator.

Translators can use these suggestions, adapt them if necessary or decide to translate from scratch.

Computer- Aided Translation

How can CAT Tools help to increase efficiency and to reduce cost thus allowing for a higher turn-around?

CAT Tools contain a translation memory (TM)

Translations are saved in

When the a sentence occurs that has already been translated similar ce) the red tion are sted to nslator.

Translators can use these suggestions, adapt them if necessary or decide to translate from scratch.

In addition to greater efficiency this also leads to increased consistency.

Computer- Aided Translation

Typical components of CAT-Tools

Translation Memory: Database in which translations are stored (typically as sentence pairs).

Termbase: Database in which terminology is stored and managed.

Editor: Writing environment to create and to edit translations.

Alignment: Application to recyle legacy data.

Filter: Tool to convert various file fomats in translatable file formats.

Project- or Workflowmanagement

CAT Tools Komponenten

Translation Memory

CAT Tools – Translation Memory

2 Definitions

"A translation memory is a text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions."
(EAGLES 1996:140)

Database in which translation units (sentence pairs) are stored. Each segment in the source language is stored with the corresponding segment in the traget language.

(across academy: http://www.across-academy.net/de/glossary-tuv.aspx)

CAT Tools – Translation Memory Systeme

Translation Memory

TM applications enable the user to save his/her translation in a database.

During translation the source text is compared with the contents of the database.

If the same sentence (or a similar one) is found in the database, the system suggests the stored translation to the translator.

If there is no match in the database (translation memory) the translator creates a new translation and saves it in the database.

CAT Tools - Segmentierung

Segmentation

- For translation the text is segmented in smaller parts (mostly sentences or paragraphs).
- Diese Einheiten heissen *Segmente*.
- The segmentation is based on a set of rules,
  - taking into account punctuation, spaces, special characters (tab, paragraph marks), capitalization and other mostly formal criteria.
- Hence the importance to structure and format the source text "correctly".

CAT-Tools Segmentation

Examples for segmentation issues: Copy & Paste from PDF

Aus vielen Bereichen der Übersetzungstätigkeit ist der Einsatz von Translation-¶
Memory-Systemen (TM-Systeme) inzwischen nicht mehr wegzudenken. Durch¶
Wiederverwendung bereits übersetzter Textpassagen trägt die Nutzung solcher¶
Systeme zu einem einheitlicheren Stil und höherer terminologischer Konsistenz¶
von Übersetzungen bei. Ob sich darüber hinaus auch die von Kunden vielfach¶
erwarteten Kosteneinsparungen und die von Berufspraktikern erhofften Effizienzgewinne¶
in ihrer Tätigkeit erzielen lassen, hängt maßgeblich davon ab, wie gut¶
der Nutzer diese komplexen Werkzeuge beherrscht.¶
Nicht wenige Kolleginnen und Kollegen mussten in den vergangenen Jahren feststellen,¶
dass die Einarbeitung in ein TM-System nach der „Trial and Error"-Methode¶
nur sehr bedingt von Erfolg gekrönt war, und nicht selten versauerte die (womöglich¶
mit hohem finanziellem Aufwand angeschaffte) Software nach einigen¶
mühevollen, zeitraubenden und frustrierenden Einarbeitungsversuchen ungenutzt¶
auf der Festplatte – nur um fortan noch durch Fehlermeldungen beim Hochfahren¶
des Computers gelegentliche „Lebenszeichen" von sich zu geben.¶

In MS Word

CAT-Tools Segmentation

Examples for segmentation issues: Copy & Paste from PDF

	Beispiel Copy aus PDF.docx		Beispiel Copy aus PDF.docx
1	Aus vielen Bereichen der Übersetzungstätigkeit ist der Einsatz von Translation-		
2	Memory-Systemen (TM-Systeme) inzwischen nicht mehr wegzudenken.		
3	Durch		
4	Wiederverwendung bereits übersetzter Textpassagen trägt die Nutzung solcher		
5	Systeme zu einem einheitlicheren Stil und höherer terminologischer Konsistenz		
6	von Übersetzungen bei.		
7	Ob sich darüber hinaus auch die von Kunden vielfach		
8	erwarteten Kosteneinsparungen und die von Berufspraktikern erhofften Effizienzgewinne		
9	in ihrer Tätigkeit erzielen lassen, hängt maßgeblich davon ab, wie gut		
10	der Nutzer diese komplexen Werkzeuge beherrscht.		
11	Nicht wenige Kolleginnen und Kollegen mussten in den vergangenen Jahren feststellen,		
12	dass die Einarbeitung in ein TM-System nach der „Trial and Error"-Methode		
13	nur sehr bedingt von Erfolg gekrönt war, und nicht selten versauerte die (womöglich		
14	mit hohem finanziellem Aufwand angeschaffte) Software nach einigen		
15	mühevollen, zeitraubenden und frustrierenden Einarbeitungsversuchen ungenutzt		
16	auf der Festplatte – nur um fortan noch durch Fehlermeldungen beim Hochfahren		
17	des Computers gelegentliche „Lebenszeichen" von sich zu geben.		

In a typical CAT Tools (SDL Trados Studio)

CAT-Tools Segmentation

Example: Law texts

Beispiel·Gesetzestext¶

[3]·Als·Ursprungsland·des·Werkes·gilt:·für·die·veröffentlichten· Werke·das·Land·der·ersten· Veröffentlichung,· selbst·wenn·es·sich·um·Werke·handelt,·die·gleichzeitig· in·mehreren· Verbandsländern·mit·gleicher·Schutzdauer· veröffentlicht· wurden;·wenn·es·sich·um·Werke·handelt,· die·gleichzeitig· in·mehreren· Verbandsländern·mit·verschiedener· Schutzdauer· veröffentlicht· wurden,·das·Land,·dessen·Gesetzgebung· die·am·wenigsten·lange·Schutzdauer· gewährt;·für·die· Werke,·die·gleichzeitig· in·einem·verbandsfremden· Land·und·in·einem·Verbandsland·veröffentlicht· wurden,·gilt·ausschliesslich·das·letztere·als·Ursprungsland.· Als·gleichzeitig· in·mehreren·Ländern· veröffentlicht· gilt·jedes·Werk,·das·innerhalb·von·dreissig· Tagen·seit·der·ersten·Veröffentlichung· in· zwei·oder·mehreren·Ländern·erschienen·ist.¶

Aus:··Berner·Übereinkunft·zum·Schutze·von·Werken·der·Literatur·und·der·Kunst·revidiert·in·Brüssel· am·26.·Juni·1948¶

http://www.admin.ch/opc/de/classified-compilation/19480180/index.html¶

In MS Word

CAT-Tools Segmentation

Example: Law texts

	Beispiel Gesetzestext.docx		Beispiel Gesetzestext.docx
1	Beispiel Gesetzestext		
2	cf 3 3 cf cf Als Ursprungsland des Werkes gilt: für die veröffentlichten Werke das Land der ersten Veröffentlichung, selbst wenn es sich um Werke handelt, die gleichzeitig in mehreren Verbandsländern mit gleicher Schutzdauer veröffentlicht wurden; wenn es sich um Werke handelt, die gleichzeitig in mehreren Verbandsländern mit verschiedener Schutzdauer veröffentlicht wurden, das Land, dessen Gesetzgebung die am wenigsten lange Schutzdauer gewährt; für die Werke, die gleichzeitig in einem verbandsfremden Land und in einem Verbandsland veröffentlicht wurden, gilt ausschliesslich das letztere als Ursprungsland. cf		
3	Als gleichzeitig in mehreren Ländern veröffentlicht gilt jedes Werk, das innerhalb von dreissig Tagen seit der ersten Veröffentlichung in zwei oder mehreren Ländern erschienen ist.		
4	Aus:		
5	Berner Übereinkunft zum Schutze von Werken der Literatur und der Kunst revidiert in Brüssel am 26.		

In a typical CAT Tools (SDL Trados Studio)

CAT-Tools Segmentation

Segmentation: Rule of thumb

- The shorter the segment the better the chance for re-use.
- The longer the segment, the more context is include --> the better the quality.

CAT Tools - Matches

Matches and types of matches

The correspondence of the segment to be translated with a segment in the translation memory is called a match.

Identical segments are called **100% matches**

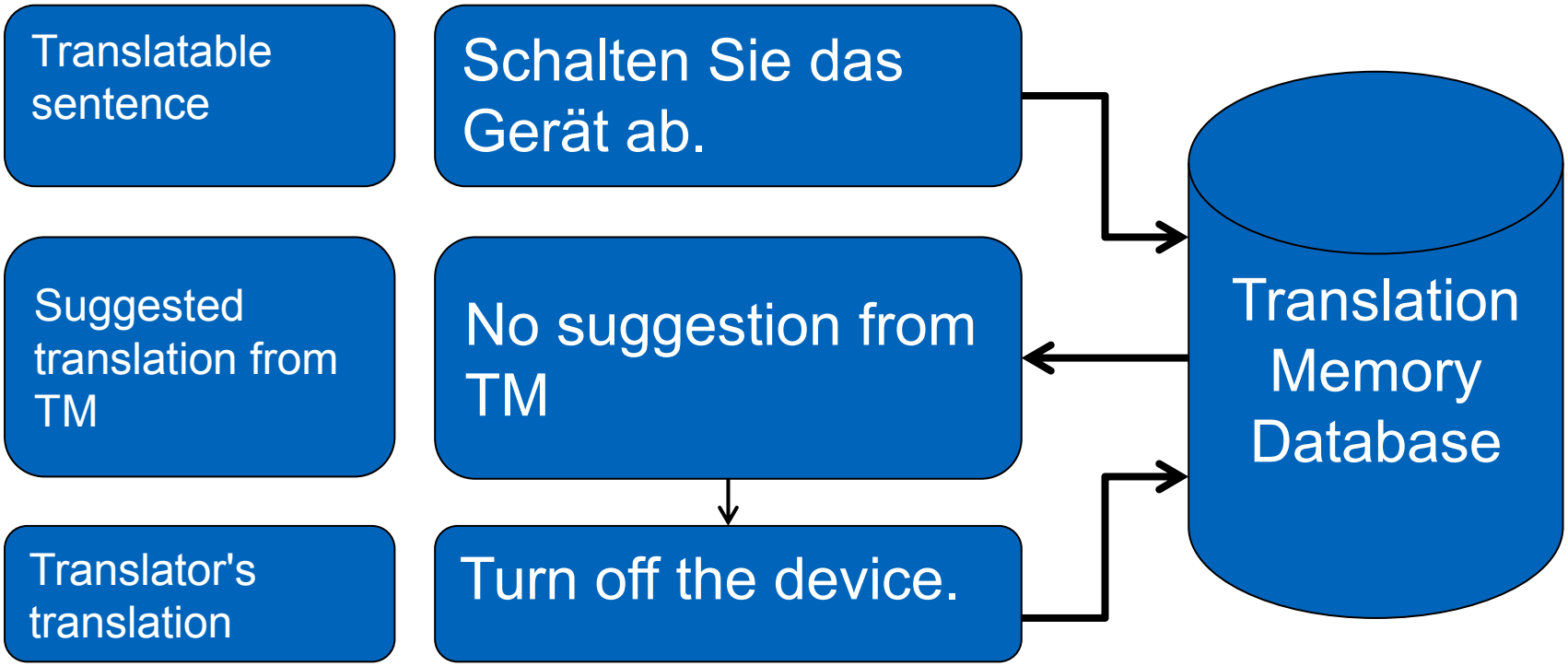
Similar segments are called **fuzzy-matches**

Segments that are not only identical but also appear in the same context are called **context-matches**

CAT Tools

Example Translation Memory

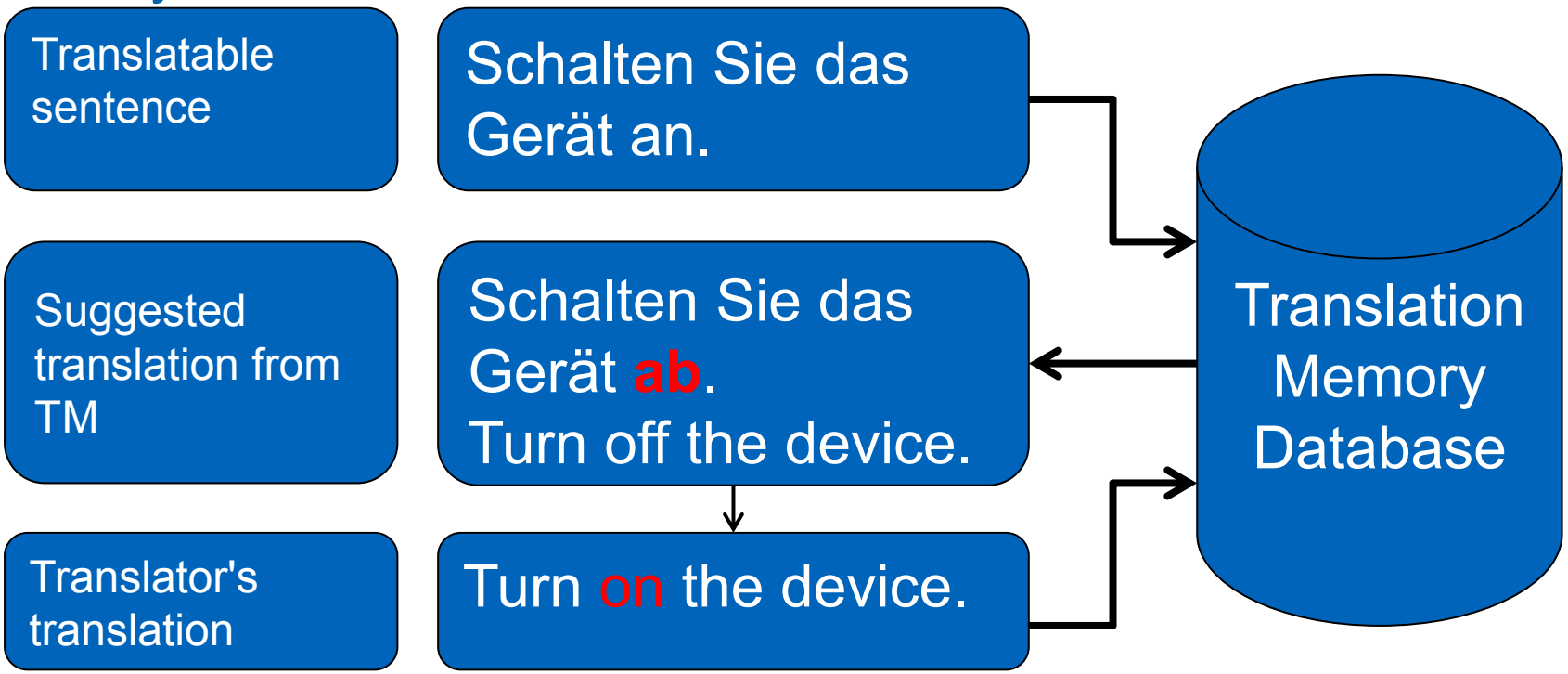
No Match



CAT Tools

Example Translation Memory

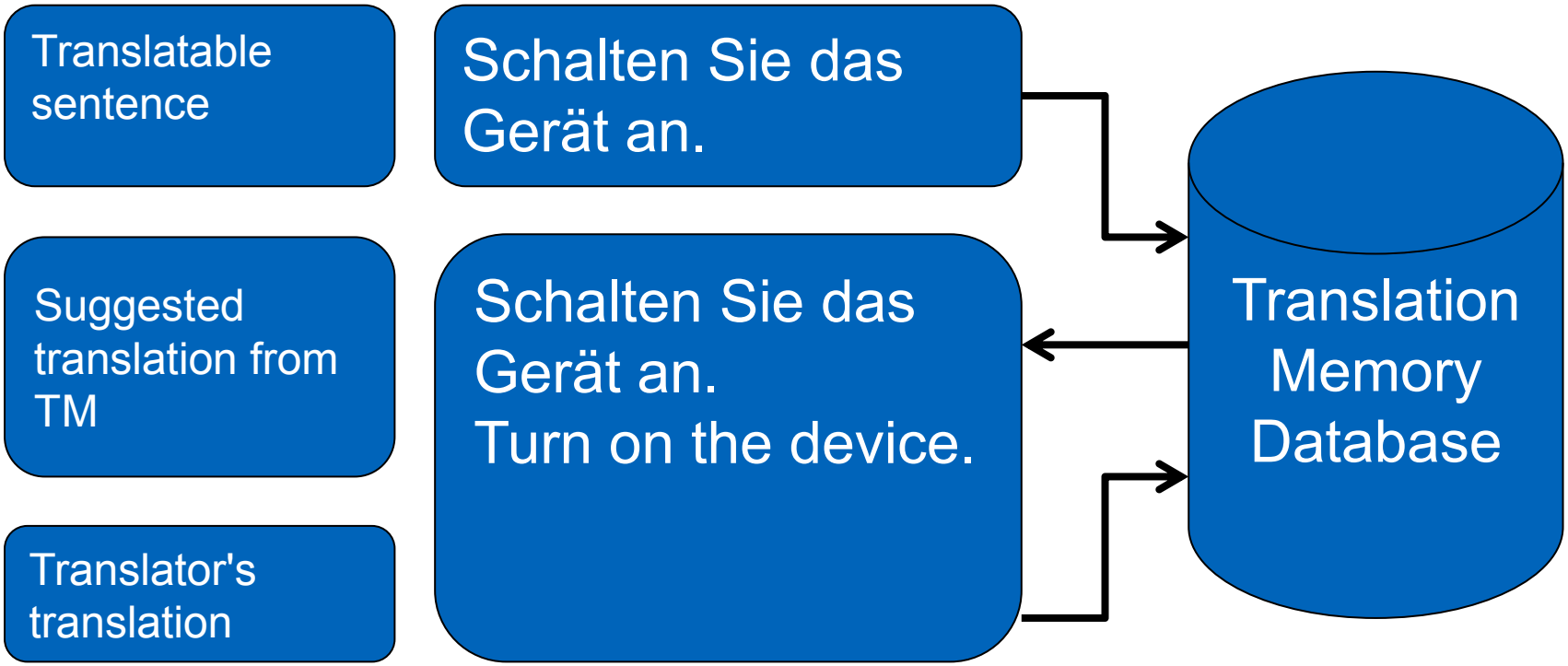
Fuzzy Match



CAT Tools

Example Translation Memory

100% Match



CAT Tools – Herausforderungen

100% Match challenge

100% Matches are not always reliable

Example:

- Do not touch the cover. It may be hot.
- Abdeckung nicht berühren. **Sie könnte heiss sein.**

- Do not touch the cable. It may be hot
- Kabel nicht berühren. **Es könnte heiss sein.**

CAT-Tools: Hands-On Exercise

Guided Exercise

- Translation of a small Microsoft Word file (To translate.docx) from English to German (recommended) using the web-based CAT-Tool MemSource
- Logon to https://cloud1.memsource.com
- Username: UNI KN X (X = 1 – 12)
- Password: UniKonstanzX (X = 1 – 12)
- Resources: Translation Memory w/ 13 Segments, termbase w/ 2 entries, Microsoft Machine translation engine

CAT-Tools: Hands-On Exercise

Selected References:

- Arnold, Doug. 2008. Machine Translation: an Introductory Guide. London: Blackwell. http://www.essex.ac.uk/linguistics/external/clmt/MTbook/
- Carstensen, Kai-Uwe, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde und Hagen Langer (eds.). 2001. *Computerlinguistik und Sprachtechnologie | Eine Einführung.* Heidelberg: Spektrum-Verlag.
- Hutchins, John. 2005. The history of machine translation in a nutshell. http://www.hutchinsweb.me.uk/Nutshell-2005.pdf
- Hutchins, John 2003. Machine translation: general overview. In: Mitkov, Ruslan (ed.) ***The Oxford Handbook of Computational Linguistics***. (Oxford: University Press, 2003), 501-511. http://www.hutchinsweb.me.uk/Mitkov-2003.pdf