



UNIVERSITÀ DEGLI STUDI DI MACERATA

Dipartimento di Studi Umanistici – Lingue, Mediazione, Storia, Lettere, Filosofia 1

*Corso di Laurea Magistrale in Lingue Moderne
per la Comunicazione e la Cooperazione Internazionale (Classe LM-38)*

Housekeeping announcements

- Info on the lecturer, e.g. (changes to) weekly office hours, etc.:
 - <http://docenti.unimc.it/docenti/federico-gaspari/>
- Info on the module, e.g. module pack, cancelled/rescheduled lectures:
 - can be accessed from personal webpage > Insegnamenti AY 2012/2013
- Slides and assignments in the module pack (please download ASAP!)
- Please sign the attendance list
- How do we structure the class? Beginning/end/break? Over to you...
- Consultation (office hour): Tuesday at 11:00 AM, Aula F Palazzo Ugolini
- All students should have access to online computers outside class



UNIVERSITÀ DEGLI STUDI DI MACERATA

Dipartimento di Studi Umanistici – Lingue, Mediazione, Storia, Lettere, Filosofia 2

*Corso di Laurea Magistrale in Lingue Moderne
per la Comunicazione e la Cooperazione Internazionale (Classe LM-38)*

- Lectures are in English (including students' questions and comments!)
- Do not be afraid to ask questions, contribute comments, thoughts, etc.
- Active participation in class + independent work at home are essential
 - this involves 2-3 hours of work outside of class every week for preparation
- In this module we are not going to work on the language as such – I assume that your English is (very) good!
- You have to keep working hard to improve your English (language tutorials are crucial for your preparation for the final written exam!)
- Essential to use a good monolingual English dictionary (also in the exam!)
- Check the module web page regularly for updates and additions (e.g. further materials for the module pack, exercises, activities, etc.)
- If you have a problem, question or doubt, do not wait and get in touch!



UNIVERSITÀ DEGLI STUDI DI MACERATA

Dipartimento di Studi Umanistici – Lingue, Mediazione, Storia, Lettere, Filosofia 3

*Corso di Laurea Magistrale in Lingue Moderne
per la Comunicazione e la Cooperazione Internazionale (Classe LM-38)*

- 6-credit module (as for module A): 10 weekly lectures of 3 hours each
- 4 main topics
 - The Internet, search engines and online resources for specialised translation
 - Machine translation (including pre-, post-editing and controlled language)
 - Overview of computer-assisted translation (CAT) tools: translation memory software and terminology management packages
 - Construction, interrogation and use of (web-derived) electronic corpora
- Books on the reading list for this module:
 - the lectures are to some extent based on these books
 - attending classes regularly and reading these sources in English helps your preparation for the final written exam in English (extra practice)



UNIVERSITÀ DEGLI STUDI DI MACERATA

Dipartimento di Studi Umanistici – Lingue, Mediazione, Storia, Lettere, Filosofia 4

*Corso di Laurea Magistrale in Lingue Moderne
per la Comunicazione e la Cooperazione Internazionale (Classe LM-38)*

- Only 1 adopted/set textbook: Somers (ed.) (2003)
 - most of this book is going to be useful (very relevant and comprehensive)
- In addition, there are recommended/suggested books
 - only some parts of these books are going to be needed/used (relevant pages/ sections are indicated in the slides for specific topics), but they are helpful sources for further study, e.g. for topics that you find particularly interesting and would like to learn more about, or if you need extra explanation...
 - Austerlühl (2001): highly accessible, but much of the content is very dated
 - Bowker (2002): particularly good for CAT tools, including translation memory
 - Bokwer & Pearson (2002) + Zanettin (2012): construction and use of corpora
 - Quah (2006): fairly wide and comprehensive scope, not too technical
 - Bersani Berselli, G. (a cura di) (2011) *Usare la traduzione automatica*. Bologna: CLUEB
 - Lecci, C. & E. Di Bello (2012) *Usare la traduzione assistita*. Bologna: CLUEB



UNIVERSITÀ DEGLI STUDI DI MACERATA

Dipartimento di Studi Umanistici – Lingue, Mediazione, Storia, Lettere, Filosofia **5**

*Corso di Laurea Magistrale in Lingue Moderne
per la Comunicazione e la Cooperazione Internazionale (Classe LM-38)*

- Preliminary information on the format of the end-of-module exam
- The official information with further details will be announced at the end of the module, this is just for you to get a rough idea of how you're going to be assessed for module B (separate exam from module A)
 - written exam (2 hours in total)
 - questions to answer from a range of options (choose 2 questions out of 4)
 - each of the available questions focuses on a different area/topic of the module
 - all questions are equally weighted and must be answered in English
 - answers are essay-type and/or require you to present an argument/opinion
 - indicative number of words is given for expected length of each answer (300)
 - only monolingual English dictionary is allowed
 - assessment is based on the correctness and completeness of the answers, taking into account the overall quality of your written English at all levels: grammar, vocabulary, syntax, clarity, fluency, etc.



Extended units of meaning

6

- “Words have meanings” vs “meanings have words”
- “Extended units of meaning” (Sinclair, 1996) at work in language
 - collocation
 - colligation
 - semantic preference
 - semantic prosody (a.k.a. discourse prosody)
- We exploit these linguistic notions to use the Internet as a translation-oriented resource (Sinclair did not connect them to the Web!)
- Concepts developed within corpus linguistics (cf. computational ling. course!)
- Empirical approach to linguistics (based on empirical observation)
- Particularly useful for specialised translators and linguists
- Focus on uncovering patterns and regularities at work in language
 - deviations, innovations, surprising and creative usage are interesting too (e.g. puns)



Collocation

7

- “You shall know a word by the company it keeps” (J.R. Firth, 1957),
i.e. you don’t know a word unless you know its behaviour and environment
- “Tendency of certain words to co-occur regularly in a given language”
(Mona Baker, 1992: 47)
- Key features of collocations
 - language-specific (collocations vary from language to language)
 - semantically arbitrary
- Collocations are not stable or fixed
 - they change diachronically (over time) in general language
 - they change in LSP vs. general language
 - they change across LSP domains



Collocation

8

- Collocations can be
 - either *unmarked*
 - or (highly) *marked*
 - and words are said to collocate strongly or weakly
- *Node*: word under scrutiny (to find its collocates)
- *Collocational range*: set of collocates (frequently co-occurring words)
- Further (suggested/optional) reading on collocation
 - Baker, Mona (1992) *In Other Words. A Coursebook on Translation*. London: Routledge. Pages 46-ff.
 - Laviosa, Sara (2008) *Linking Wor(l)ds. Lexis and Grammar for Translation*. Napoli: Liguori Editore. 2nd edition. Pages 15-ff.



Colligation

9

- Relation between a pair of grammatical categories or a pairing of lexis and grammar (Stubbs, 2001: 65)
- “achieving aims”, “aims being achieved”, “achievable aims”, “the achievement of an aim” are all standard/acceptable colligations
- “drink water”, “the water is (not) drinkable” both fine
- We say “break the law(s)”, but not “the law(s) is/are (not) breakable”
- Corresponding collocations and colligations in Italian for “break the law”?



Semantic preference

10

- Relation between a lemma and a set of semantically related words (Stubbs, 2001: 65)
- *Lemma*: base form (lexeme) or dictionary entry of a word
- “Commit” is used with a group of semantically similar words, e.g. “murder”, “crime”, “suicide” (cf. Italian “commettere”)



Semantic prosody (discourse prosody)

11

- “Feeling” or “aura” that is evoked by using certain words (reinforced by collocates, due to co-selectional implications and restrictions)
- Usually this feeling is “positive or negative”, but not only/always
- E.g. in Italian “Il paziente versa in X condizioni” (X positive or negative?)
- “Provide” tends to occur with words denoting things which are desirable, necessary or good, such as “information”, “service(s)”, “support”, “help”, “money”, “protection”, “food”, “care” (cf. Italian “fornire” and “elargire”)
- There are borderline/ambivalent cases (e.g. the Italian verb “mietere”), or words with a (seemingly) neutral or unspecified semantic prosody
- Theory of **Lexical Priming** (Hoey, 2005): words are primed to occur in certain environments
- *Lexical profile* is a combination of all the above information, a sort of ID of a word, characterising its typical behaviour and environment



Tips and tricks for effective Internet searches

12

- Note that searches are not case sensitive
 - **“AAAA BBBB CCCC”**
returns hits containing the words (as many as you want) in this sequence
 - **“XXXX * YYYY ZZZZ”** (the * is known as *wildcard*)
returns hits containing the sequence of words XXXX, something unspecified in the middle – one or more word(s) – and then YYYY ZZZZ
 - **+KKKK**
returns hits containing this word (+ is not necessary for a single word)
 - **-JJJJ**
returns hits that do not contain this word
- You can use this syntax in combination to perform advanced textual searches, e.g.: **+“XXXX * YYYY ZZZZ” +KKKK -JJJJ**
(leave one blank space between the search criteria)



- Translators can use search engines to find/verify two kinds of things
 - facts, pieces of information, knowledge (as an “encyclopaedia”)
 - words, terms, collocations, phrases, expressions, instances of language use
- We can use the Internet (via a search engine) to look for patterns and regularities in (our native, foreign, source and/or target) language
- We extract examples of real language in use, which we can analyse to infer repeated, common or typical linguistic behaviour (e.g. collocates)
- My experience as someone who uses (and learns) English every day is that dictionaries are not sufficient (although you must learn to use a monolingual English dictionary, which is crucial for the written exam!)
- The point is to devise strategies to mine the Web for (useful) linguistic information from which we can learn or discover something interesting



- *Collocational patterns* of node word “holiday(s)”
- *Collocational patterns* of nodes “issue” and “issues” (as nouns vs. verbs)
- *Colligations* of “break the rule(s)”, “the rule(s) is/are (not) breakable”
- *Semantic preference* of “momentous”
- *Semantic prosody* of “cause” (as a verb)
- After looking at repeated uses in context, we can establish with a good degree of confidence the *lexical profile* of several words and lexical items, understanding how they are used in *extended units of meaning*, observing the environment in which they are *primed* to occur



- ADJ-NOUN collocational patterns for a given NOUN (cf. ex. 1, p. 203)
- How can we translate into English “un’ ampia / una vasta gamma di...”
- Compare the collocates of two similar words (nouns)
 - RANGE
 - ARRAY
- The * is known as *wildcard*
- What patterns do we search for? Some possibilities to start with
 - “a * NOUN”
 - “an * NOUN”
 - “this * NOUN”
 - “such a(n) * NOUN”
 - “have a(n) * NOUN”
 - “has a(n) * NOUN”
 - “our * NOUN”
 - “their * NOUN” etc. etc.



• **ARRAY**

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

• **RANGE**

- A
- B
- C
- D
- E
- F
- G
- H
- I
- J



***Advanced Internet search techniques:
collocations in general language vs. LSP***

17

- Strong collocates of the Italian node word “naso” in general language
- What are the members of the collocational range of “naso” in Italian?
- Differences in Italian between (from Taylor, 1998: 61):
 - “pressione alta” = “high (blood) pressure” [medical]
 - “alta pressione” = “(banks of) high pressure” [meteorological]
- Does “naso ampio” seem an unmarked (typical) collocation to you?
- Search for the exact phrase “naso ampio” on google.it
- Does “naso ampio” seem to be an established/frequent collocation?
- It is a frequently attested collocation used by specialists of wine tasting!
- Look at a few examples and try to understand what “naso ampio” means
- Check “ampio naso” (reverse sequence) – is it used in similar ways?



***Preferred context-dependent uses and
translating collocations into English***

18

- Are there any differences in the use of these two similar adjectives?
 - powerful
 - potent
 - search on google.co.uk for phrases such as “is a ADJ”, “this/these ADJ”, “very ADJ”, “extremely ADJ”, etc.
 - pay special attention to the sources and surrounding words
 - do these two adjectives seem to be used interchangeably?
 - which one is more frequently found in medicine and/or chemistry?
- Are there any differences in the use of these two similar phrases?
 - “homely environment”
 - “homely atmosphere”
 - search for these two phrases separately on google.co.uk
 - with what places or contexts is each of them more commonly associated?
- Do you know what “al fresco” means in English? Find out in what context this phrase is used in English, searching for it on google.co.uk



*Seemingly highly marked (but absolutely unmarked!)
collocations in standard English*

19

- Words with an adjectival function that collocate strongly with “change”
 - search for “a * change” and “there has been a * change” on google.co.uk
 - do you find any surprising/unexpected collocates replacing the wildcard?
 - collocations are semantically arbitrary! (cf. Italian “*differenza abissale*”)
- How can you qualify/describe idiomatically “a lot of rain” in English?
 - search for “rain was coming down (in)” on google.co.uk
 - do you find any surprising/unexpected collocates on the right?
 - we were perhaps expecting “torrents” and “buckets”...
 - ... but what about “stair rods” and “sheets”?!? They sound quite weird!
- Is this so surprising or unusual as it looks at first sight?
 - what is the unmarked collocation in Italian to describe “areas of fog”?
 - e.g. “c’è nebbia a X” “X di nebbia sull’autostrada A14 verso Bologna”
 - X = “banchi”!! What’s the reason for this puzzling collocational behaviour?



*Investigating marked collocational uses
in L2 writing or translations*

20

- Think of the Italian adjective “suggestivo”, typical of tourism texts, e.g.
 - “È situato a pochi minuti dalla stazione Termini ma in un angolo **suggestivo** e tranquillo della città”
 - “Portofino è una località molto esclusiva immersa in un paesaggio **suggestivo** e in una natura incontaminata”
- The problem is how to translate the adjective “suggestivo” into English
- Under the entry for “suggestivo”, most bilingual Ita-Eng dictionaries give e.g. “striking, attractive, evocative, charming, suggestive”, etc.
- So it seems a safe bet to translate into English with “suggestive”, right?
- First let’s check “in a suggestive” and “is a suggestive” on google.co.uk
- Then let’s check “in a suggestive” and “is a suggestive” on google.it



Investigating marked collocational uses in L2 writing or translations

21

- Finally, let's double-check how widely used some phrases are on
 - google.co.uk (.uk pages only) vs. google.it (.it pages only)to study the native vs. non-native collocational behaviour of the node word
- Search for these exact phrases on the language-specific search engines:
 - “suggestive environment”
 - “suggestive setting(s)”
 - “suggestive position”
 - “suggestive scenery”
 - “suggestive landscape”
- Are these phrases more often used in .it or .uk webpages?
- When the phrases are attested on .uk webpages, what do they refer to?
- What does this situation suggest? (*pun not intended...*)



Summary on using the Internet for specialised translation

22

- What the Internet (and Google) can and cannot do for translators
- We have used the Web
 - to look for linguistic information (patterns), mono/biling. dictionary-style
 - to check/verify culture-bound facts, encyclopaedia-style
 - for language learning/study
 - for specialised translation, as a quick-reference tool for technical translators
 - plus for a range of language professionals, e.g. interpreters, revisers, etc.
- We have used the Web to solve a variety of different problems
 - linguistic: understand meaning, uses and associations of words/phrases
 - translational: attested loans, confirm hypotheses on candidate equivalents
 - cultural: items/references familiar in SC, unknown to translator and TC



Conclusion on using the Internet for specialised translation

23

- Main advantages
 - massive amount of text and multi-source information can be searched
 - content is constantly “refreshed” (i.e. updated and extended)
 - lots of sources, text types and domains/topics are represented
 - many languages (English is dominant)
 - replicable search techniques across (your working/target) languages
 - it’s available at any time, at virtually no cost!



Conclusion on using the Internet for specialised translation

24

- Main disadvantages and problems
 - need to differentiate good/reliable sources from questionable information
 - for facts (limited control over user-generated content like Wikipedia)
 - for linguistic usage (badly translated, non-native texts, poor authors)
 - it may be difficult to identify differences between expert/non-expert use
 - data/results still need to be interpreted (hard work, but gets faster!)
 - Google focuses on content/information, rather than linguistic forms
 - it is a commercial search engine: we have no control over it!
 - its internal search algorithms are only partially transparent
 - e.g. difficult to find the collocates of the word “gore” (in “blood” sense), as there is a lot of noise, coming from “Al Gore”, “Gore-Tex”, etc.



Conclusion on using the Internet for specialised translation

25

- Main disadvantages and problems (continued)
 - particularly relevant to us
 - no possible/meaningful sorting of hits/results (esp. L/R-hand collocates)
 - e.g. alphabetical sorting of collocates, from least to most frequent, etc.
 - punctuation and upper case (capitals) are ignored, e.g. “aids” vs. “AIDS”
 - impossible to search parts of words, e.g. start with “geo...”, end in “-itis”
 - no (or limited, anyway) lemmatised stemmed searches
 - no POS-sensitive searches (e.g. search for adjective vs. noun vs. verb)
 - need workarounds to focus on specific patterns/phrases (filling in slots)
 - data on which searches are performed is unstable/changes
 - word counts are not reliable (try “we asked for your support”)
 - difficult to compare frequency to verify translation hypotheses
 - need to run several parallel searches and compare between their results



Readings

26

- Four chapters from Austermühl, F. (2001) *Electronic Tools for Translators*.
Manchester, St. Jerome Publishing, i.e.

- + 2 “Translator-client communication and information transfer”, pp. 18-36
- + 3 “Translation and the Internet”, pp. 38-50
- + 4 “Searching the web”, pp. 52-67
- + 5 “Translation resources on the worldwide web”, pp. 68-89

The information contained in these chapters is useful background reading, as the concepts and issues are explained in a highly accessible style. However, be aware that the book was published in 2001, so most of the information and resources that are discussed in it regarding the Internet and translation are now very dated. As a result, these chapters should be supplemented by the handouts, exercises and take-home assignments that were used in class and for homework.



***(Fun) Homework set
in preparation for next class***

27

- Try out any of the four free online machine translation services listed below
- You can translate your own sentences and short texts (typed in or copied & pasted) or provide Internet addresses (URLs) to translate entire webpages
- Translate any text or webpage that you like between English and Italian (in both directions); feel free to test any other language combination as well
- Translate a random set of different sentences, texts or webpages until you get an idea of the quality of the translations. How would you judge them in general? Are there any obvious errors? Are the target texts understandable?
 - <http://babelfish.yahoo.com>
 - <http://www.freetranslation.com>
 - http://www.google.com/language_tools
 - <http://www.microsofttranslator.com>



***Introduction to
machine translation (MT)***

28

- Overview
 - Definition and key terms
 - Brief outline of “historical” origins and some recent developments
 - Main architectures of MT systems (rule-based vs. statistical approaches)
 - Why is MT so difficult? Or why is translation difficult for computers?
 - Some linguistic phenomena that are particularly difficult for MT
 - Forms of human intervention in MT
 - Restrictions to the use of MT



Machine translation (MT): definition and key terms

29

- Definition of machine translation:

“computerised systems responsible for the production of translations from one natural language into another, with or without human assistance”
(Hutchins & Somers, 1992: 3)

- Human intervention is not necessarily excluded, but if it does occur it is subordinated to the prevailing action of the computer

- Some key terms:

- **MT system / engine / service** = the software that produces the translation
- **input** = the source text (i.e. original that we are trying to translate)
- **[raw] output** = [unedited] target text (i.e. the translation that we obtain)



Machine translation (MT): brief outline of “historical” origins and some recent developments

30

- MT was one of the first non-numerical applications of computers
- The idea of using the computer to translate was put forward in 1949
 - at that time computers were very different from the current ones!!
 - idea formulated by W. Weaver in a “memorandum” to other scientists
 - starting point were the advances in cryptography during World War II
- Declared aim of initial research on MT in the 1950s and 1960s
 - creating systems that would be capable of offering fully automatic high-quality machine translation for unrestricted texts (in any domain)

(FAHQMT-UT)



Machine translation (MT): 31
brief outline of “historical” origins and some recent developments

- At the beginning a lot of activity in the USA, later also in the USSR
- Efforts in the US focused on RU>EN (monodirectional!) MT systems
- First demonstration in 1954 at Georgetown University (IBM involved)
 - individual pre-determined sentences translated from Russian into English
 - lexicon (~vocabulary) with 250 items, first generation (direct approach)
- ALPAC Report in 1966 (a committee of US experts - established 1964)
 - disappointing results viz. massive federal funding, stop to public money
 - insufficient need for large-scale translation
 - MT slower, less accurate and more costly than human translation (HT)
 - lack of prospects for an immediate or short-term breakthrough
 - need to invest in basic research (comp. linguistics) and translator training



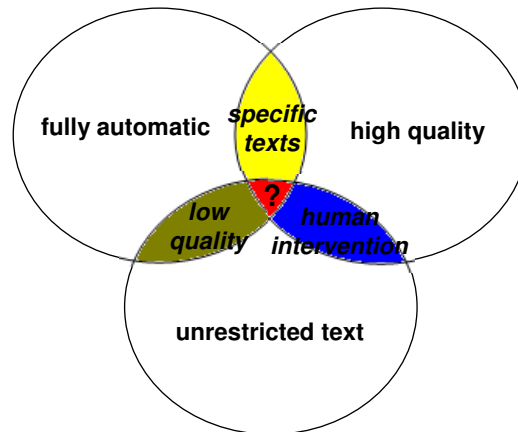
Machine translation (MT): 32
brief outline of “historical” origins and some recent developments

- Today most (reasonable) people agree that (A) fully automatic (B) high-quality MT of (C) unrestricted texts is not possible
- You have to make a compromise and sacrifice at least one of these three requirements (can't have all of them at the same time):
 - A - give up full automation (i.e. involve humans in the MT process, e.g. pre-edit the input, post-edit the output, or have an interactive system)
 - B - accept less-than-perfect translation quality (very poor, most of the time)
 - C - tailor MT system to translate only texts in a well-defined limited domain (cf. statistical MT later on)



Machine translation (MT):
brief outline of “historical” origins and some recent developments

33



Machine translation (MT):
brief outline of “historical” origins and some recent developments

34

- MT today is heavily used on the Web (thanks esp. to free online services)
- Who has used them? What did you translate? What do you think?
rubbish bad so-so good brilliant
- Were they fast? Think in terms of quality/speed/cost!
- Babel Fish launched in December 1997 by the search engine Alta Vista in partnership with the well-known MT company Systran
 - Babel Fish (*now Yahoo! offers 34 language pairs, initially 10*)
 - FreeTranslation (*13/18 language pairs*)
 - Microsoft Translator (*22 SLs and TLs*)
 - Google Language Tools (*among 50 languages in all combinations!*)
- Many problems, mistakes and limitations, but Internet users seem to be tolerant: an imperfect translation is better than (understanding) nothing!

FreeTranslation:

3.4 million translations per day,
i.e. roughly 50 million SL words
(September 2006)



Umberto Eco on translation
(in general, i.e. on human translation, not MT)

35

“Una traduzione non è una fonte: è una protesi, come la dentiera, o gli occhiali, un mezzo per raggiungere in modo limitato qualche cosa che si trova fuori dalla mia portata.”

Eco, U. (1977) *Come si fa una tesi di laurea. Le materie umanistiche*. Milano: Bompiani



Machine translation (MT):
brief outline of “historical” origins and some recent developments

36

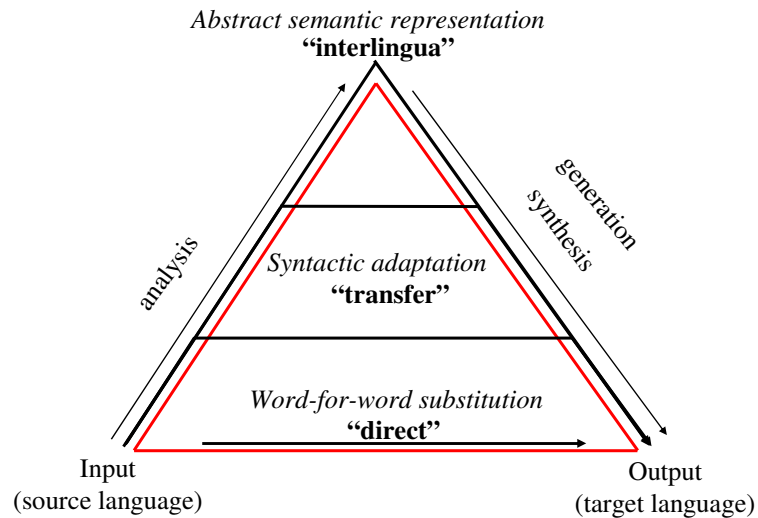
- But who uses MT “seriously” today? Why? How?
- “Serious” users motivated by the need to optimise quality/speed/cost of translations, in relation to the intended use/duration of the target texts
- Suppose you translate an easy general-language text L2 > L1 (EN > IT)
 - How fast can you go? How many words do you translate per day / hour?
- Proprietary (in-house) MT systems, developed internally, used with a licence or specifically built, usually with technical/spec. dictionaries
 - Pan-American Health Organization (PAHO): ENGSPAN, SPANAM
 - 42,000 words / hour = 700 words / minute
 - EU Commission: EC Systran, 28 language combinations (phased out)
 - 2,000 pages / hour



Machine translation (MT): main architectures of MT systems

37

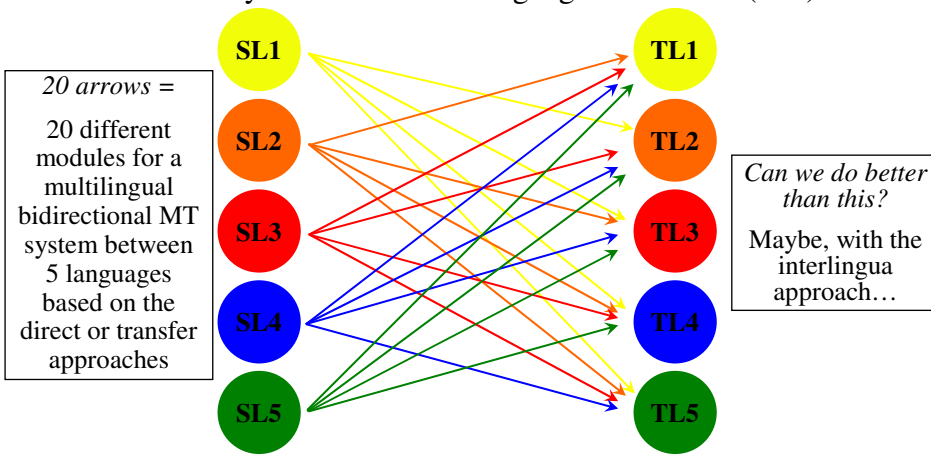
- “Rule-based” (classic) approaches to MT: the Vauquois triangle



Machine translation (MT): main architectures of MT systems

38

- You need separate “modules” for the direct and transfer approaches
- Language combinations are not necessarily symmetrical pairs
- A bidirectional system between N languages needs $N \times (N-1)$ modules

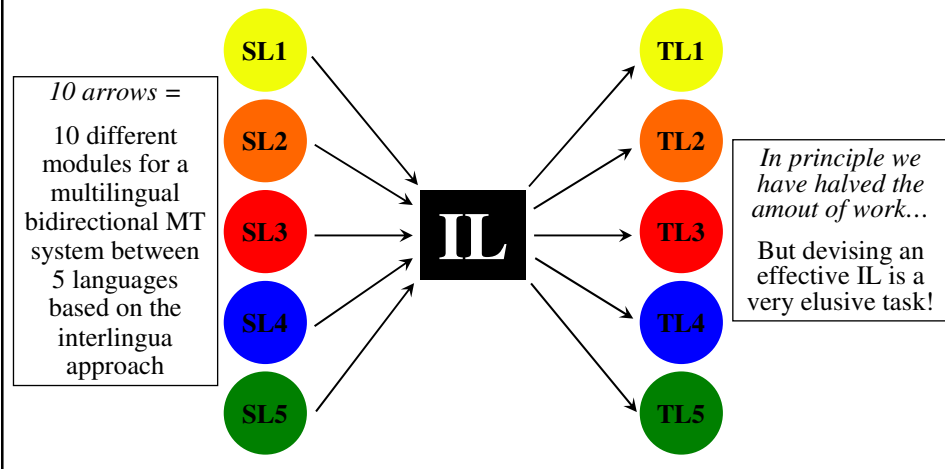




Machine translation (MT): main architectures of MT systems

39

- The interlingua (IL) consists of a set of abstract language-independent semantic representations – based on natural language or any other code
- This approach does offer some advantages for MT system design



Machine translation (MT): main architectures of MT systems

40

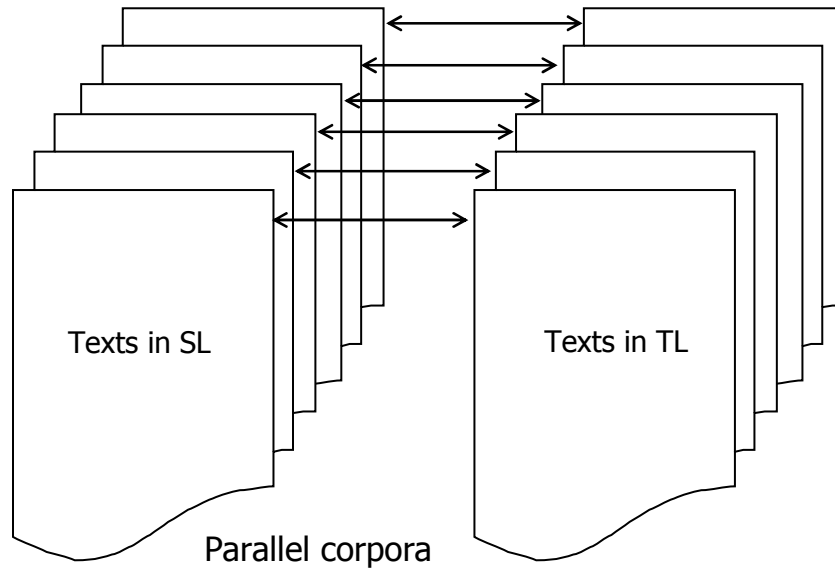
- Other new approaches to MT system design emerged in the 1990s (IBM)
- The idea is to do away with all linguistic rules, favouring empirical “statistical” and “example-based” approaches (statistical MT = SMT)
- Massive availability of electronic (translated) texts and parallel corpora
- Algorithms trained to detect and extract translational patterns on very large datasets of SL/TL texts contained in parallel corpora
- Usually based on tri-grams, accompanied by TL model (to limit overgeneration of TL output) – more a recombination than translation
- Problem of granularity and boundary friction – which is a “good” unit?
- These data-driven SMT systems perform well on new input similar to the texts on which their algorithms have been trained and developed
- However, it’s difficult to implement the initial radical idea of totally avoiding rules, going for a purely/strictly statistical data-driven approach
- Possibility of “hybrid” systems to varying degrees (stats + *some* rules)



Machine translation (MT): main architectures of MT systems

41

- More recent (statistical) approaches to MT (example-based MT)



Sentence-aligned parallel texts

42

Texts in EN

Texts in IT

The red house is big.	←→	La casa rossa è grande.
This is my new house.	←→	Questa è la mia nuova casa.
She lives in a big house.	←→	Lei vive in una casa grande.
I bought a new house.	←→	Ho comprato una nuova casa.
This house is very expensive.	←→	Questa casa costa molto.
This house is very big.	←→	Questa casa è molto grande.



Texts in EN

The red house is big.

This is my new house.

She lives in a big house.

I bought a new house.

This house is very expensive.

This house is very big.

Texts in IT

43

La casa rossa è grande.

Questa è la mia nuova casa.

Lei vive in una casa grande.

Ho comprato una nuova casa.

Questa casa costa molto.

Questa casa è molto grande.



Texts in EN

The red house is big.

This is my new house.

She lives in a big house.

I bought a new house.

This house is very expensive.

This house is very big.

Texts in IT

44

La casa rossa è grande.

Questa è la mia nuova casa.

Lei vive in una casa grande.

Ho comprato una nuova casa.

Questa casa costa molto. □

Questa casa è molto grande.

3/4



Texts in EN

Texts in IT

45

The red house is big.

2 / 2

This is my new house.

She lives in a big house.

I bought a new house.

This house is very expensive.

2 / 2

This house is very big.

La casa rossa è grande.

Questa è la mia nuova casa.

Lei vive in una casa grande.

Ho comprato una nuova casa.

Questa casa costa molto.

Questa casa è molto grande.



Texts in EN

Texts in IT

46

The red house is big.

This is my new house.

She lives in a big house.

I bought a new house.

This house is very expensive.

This house is very big.

?

La casa rossa è grande.

Questa è la mia nuova casa.

Lei vive in una casa grande.

Ho comprato una nuova casa.

Questa casa costa molto.

Questa casa è molto grande.



Texts in EN

The red house is big.

This is my new house.

She lives in a big house.

I bought a new house.

This house is very expensive.

This house is very big.

?

Texts in IT

47

La casa rossa è grande.

Questa è la mia nuova casa.

Lei vive in una casa grande.

Ho comprato una nuova casa.

Questa casa costa molto.

Questa casa è molto grande.

- Identified translational correspondences can be reversed SL <> TL



Potential translational correspondences found

Similarities with translation memory software (CAT)

48

Probabilities that X in SL corresponds to Y in TL estimated

Possible (hypothetical) translations of fragments generated

Check with a statistical model of the sole target language (only plausible translations retained)

Final target text assembled

NB: More or less radical and orthodox implementations of systems following a statistical approach are possible: by adding some linguistic rules you can obtain a hybrid MT system

I can add explicit grammatical or syntactic rules (e.g. grammatical ending agreement, adj-noun order in TL, etc.)



Machine translation (MT):

49

why is MT so difficult? Or why is translation difficult for computers?

- So why is translation difficult for computers?
 - Some blame the computer's lack of "real world knowledge"
 - Focus on potential translation problems for EN-IT (with a computer!!)
 - A simple example: lexical gaps and lexical asymmetries (concrete nouns)
 - legno / bosco / foresta in IT (+ EN, FR, DE and your other languages...)

<i>legno</i>	<i>bosco</i>	<i>foresta</i>	IT
<i>wood</i>		<i>forest</i>	EN
<i>bois</i>		<i>forêt</i>	FR
<i>Holz</i>	<i>Wald</i>		DE



Machine translation (MT):

50

why is MT so difficult? Or why is translation difficult for computers?

- Partly because the translation often depends on the context / situation, which the computer is not able to take into account

"The ball is in your court"



"Il pallone è nella vostra metà campo"
(the manager to the players)



"Il ballo è nella vostra corte"
(the chamberlain to the king)



Machine translation (MT):

51

why is MT so difficult? Or why is translation difficult for computers?

- Scope ambiguity (does it affect / can you preserve it in your TL?):
 - a) Old men and schoolgirls were taken to hospital
 - b) Old men and women were taken to hospital
 - c) Pregnant women and priests were taken to hospital
- Structural ambiguity of prepositional phrases (does it affect your TL?):
 - d) I saw John on the hill with my dog
 - e) I saw John on the hill with my eyes
- Naturalness of translated collocations
 - EN>IT
 - f) “pay a visit” (“pagare una visita”?)
 - g) “brush your teeth” (“spazzola [i] tuoi denti”?)
 - IT>EN
 - h) “fare i compiti” (“do / make the homework”?)
 - i) “ridente cittadina” (“laughing small town”?)

+ idiomatic
expressions

+ proper names:
° *George Bush*
° *Gordon Brown*
° *Tiger Woods*



Machine translation (MT):

52

why is MT so difficult? Or why is translation difficult for computers?

- Lexical ambiguities (gramm. category <-> meaning <-> translation)
for example, in EN: *control, bear, can, match, marks, light*
 - j) My team was eliminated in the first round (*Noun: girone*)
 - k) The cowboy started to round up the cattle (*Verb: radunare*)
 - l) We can use the round table for dinner (*Adjective: rotondo*)
 - m) Maggie is going on a cruise round the world (*Preposition: intorno al*)
- These sentences are ambiguous and very complex (for MT!):
 - n) Time flies like an arrow
 - o) Gas pump prices rose last time oil stocks fell



Machine translation (MT): 53
some linguistic phenomena that are particularly difficult for MT

- 1) The chimp eats the banana because it is greedy.
- 2) The chimp eats the banana because it is ripe.
- 3) The chimp eats the banana because it is lunchtime.

• The case / example of pronominal anaphora (resolution), difficult for MT



Machine translation (MT): 54
forms of human intervention in MT (the case of pre-editing)

- 1) The chimp eats the banana because ~~it~~ is greedy.
1a) The chimp eats the banana. **The chimp** is greedy.
1b) The **greedy** chimp eats the banana.
- 2) The chimp eats the banana because ~~it~~ is ripe.
2a) The chimp eats the banana. **The banana** is ripe.
2b) The chimp eats the **ripe** banana.
- 3)
The chimp eats the banana because ~~it~~ is lunchtime.
3a) **It is lunchtime and** the chimp eats the banana.

• Example of pre-editing: simplifying the input (eliminating anaphoras)



Machine translation (MT): restrictions to the use of MT

55

- Input must be (converted into) electronic format, e.g. through OCR
- Correct formatting and layout of the input are very important
 - e.g. spaces and hard returns should be only where required
 - the word “e r r o r” (spaced letters) would not be recognised / translated
 - spelling and typos are crucial (suppose the input is a gardening manual)
 - “Water the fowers every day” (is “to fow” a verb? Cf. “towers”)
 - “Water the pants every day” (“pants” is another English word!)
- Anybody would understand these banal mistakes, but not an MT system!
- Limited availability of language combinations (improving with SMT)
 - coverage mostly limited to “usual” big languages with commercial interest
- Structural and stylistic features of input (e.g. text type) – is it worth it?



References and readings (textbooks)

56

- Six chapters from Somers, H. (ed.) (2003) *Computers and Translation: A Translator's Guide*. Amsterdam and Philadelphia, John Benjamins, i.e.
 - + 8 (D. Arnold): “Why translation is difficult for computers”, pages 119-142
 - + 9 (P. Bennett): “The relevance of linguistics for mach. transl.”, pages 143-160
 - + 10 (J. Hutchins): “Commercial systems: The state of the art”, pages 161-174
 - + 11 (S. Bennett & L. Gerber): “Inside commercial mach. transl.”, pages 175-190
 - + 12 (J. Yang & E. Lange): “Going live on the internet”, pages 191-210
 - + 13 (J.S. White): “How to evaluate machine translation”, pages 211-244
- One chapter from Austermühl, F. (2001) *Electronic Tools for Translators*. Manchester, St. Jerome Publishing, i.e.
 - + 10 “A translator's sword of Damocles? An intro. to mach. transl.”, pp. 153-176
- Two chapters from Quah, C.K. (2006) *Translation and Technology*. Basingstoke, Palgrave MacMillan, i.e.
 - + 2 “Translation studies and translation technology”, pp. 22-56
 - + 3 “Machine translation systems”, pp. 57-92



Further optional readings (including online sources) **57**

- Gaspari, F. (2011) “Introduzione alla traduzione automatica”. Bersani Berselli, G. (a cura di) *Usare la traduzione automatica*. Bologna: CLUEB. Capitolo 1 (pp. 13-31)
- Hutchins, J. (1986) *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood. Available online at <http://www.hutchinsweb.me.uk/PPF-TOC.htm> (various chapters, which can be downloaded, provide further information on the topics discussed in the slides)
- Hutchins, W.J. & H.L. Somers (1992) *An Introduction to Machine Translation*. London: Academic Press. Available online at <http://www.hutchinsweb.me.uk/IntroMT-TOC.htm> (various chapters, which can be downloaded, provide further information on the topics discussed in the slides)
- Arnold, D.J., L. Balkan, S. Meijer, R. Lee Humphreys & L. Sadler (1994) *Machine Translation: an Introductory Guide*. London: Blackwells-NCC. Available online at <http://www.essex.ac.uk/linguistics/clmt/MTBook> (various chapters, which can be downloaded, provide further information on the topics discussed in the slides)
- Information and downloadable articles on machine translation:
 - *Machine Translation Archive* <http://www.mt-archive.info>
 - *Publications by J. Hutchins (MT history)* <http://www.hutchinsweb.me.uk>
 - *European Association for Machine Translation* <http://www.eamt.org>



MT post-editing: introduction

58

- We have seen simple examples of pre-editing (for pronominal anaphora)
 - Now let's look at post-editing (revising the raw output)
 - new skill that is acquired with experience, different from translation
 - in this scenario one has to balance and optimise quality-speed-cost, in relation to the intended use/duration of the translation
- length of use of the document
 - needs and expectations of the end user(s)
 - ability of the readers/addressees to make use of the doc.



- type, length and “visibility” of the document
 - turnaround time
 - available and viable options



MT post-editing

59

- The aim of post-editing is to make the revised output usable or understandable, with the least possible effort (quickly)
- The priority is to save time and money
- The extent and the accuracy of post-editing are negotiated/specified on a case by case basis, depending on the needs and requirements
- Different “types” and levels of post-editing (in companies, organisations):
 - no post-editing
 - internal circulation, almost never external publication
 - minimum post-editing
 - internal circulation, rarely external publication
 - full/complete post-editing (but... is it worth it?)
 - very rarely internal circulation, mostly external publication



MT post-editing and the post-editor

60

“This question [that has never really been touched upon before in the field of traditional translation] concerns the acceptance and use of half-finished texts. Within the [human translation] profession, creating half-finished texts is a non-issue because producing a partially completed translated text is not something that human translators do.”

(Allen, 2003: 297-298, my emphasis)

- Crucial question: who are the post-editors?





The skills of a post-editor

61

- Different from revising translations by colleagues ($\neq$ mistakes/corrections)
- Little training available (mostly acquired on the job)
- Excellent word-processing and editing skills
- Ability to work and make corrections directly on screen
- General knowledge of the problems and challenges faced by MT
- Specific knowledge of the particular MT system that is used
- Knowledge of SL and TL (at what level? It depends...)
- Quick in making decisions as to what and how to correct
- Ability to always balance effort/quality/time trade-off
- Ability to adapt to the different specifications required for each job
- Different from working in a CAT environment (more on this soon!)



Examples: PAHO and the EU Commission

62

- Remember the translation speed of their proprietary MT systems
- Both these institutions rely on post-editing (PE) for certain translations
- Aims and level of PE (minimum/full-complete) are decided specifically for each individual case, depending on the circumstances
- Factors to be considered (prioritised)
 - PE is there to save time and money (quality is less relevant)
 - understandability and correctness of general meaning are key
- Factors to be ignored (irrelevant in PE)
 - any detail or nuance (of information, of meaning, style, register, etc.)
 - elegance, fluency, naturalness of expression, etc.



Post-editing service at the the EU Commission

63

“A seminar on MT forms part of the training programme for new [EU] Commission officials. The MT system, its potential and limits, is presented to them together with specific guidelines about how to achieve the best possible results.”

(Petrits et al., 2001)

- The degree of development achieved for a language combination within the MT system (EC Systran) determines the need for and type/level of PE
- PE is regarded as an additional product/service on top of MT provision



Post-editing service at the the EU Commission

64

- The Rapid Post-Editing Service normally provides the revision of the raw output of the MT system EC Systran within 48 hours of the request
- It is useful in certain circumstances, for example when the “standard” professional (human) translation service is overloaded or would not be able to meet a tight deadline
- Decision whether to require PE applied to raw MT output rests with the user/requester (usually an EU official)
- The PE service helps to save time and money
 - Feedback is given to the MT developers on frequent problems/mistakes

Wagner (1985), Senez (1998a), Senez (1998b)



- Internal guidelines are there to ensure the quality of the translations, prevent problems/abuse and monitor the productivity of PE service
- Disclaimer inserted in raw MT output, to avoid unwanted dissemination

!!! RAW MT OUTPUT !!!

- The Rapid PE Service is entirely administered internally, but carried out by external freelancers (plus a few in-house employees)
- Need/type of PE changing with introduction of statistical MT in the EU
- Fee paid to external freelancers for Rapid PE Service
 - varies depending on the language combination (different MT quality)
 - set in proportion to the fee paid for the whole translated page (from scratch)
 - on average PE is paid roughly 50% of the “real/proper” translation



Translating for a Multilingual Community



Machine translation

The Commission has been using machine translation (MT) since 1976. The MT service offers 18 operational language pairs and can produce 2 000 pages of raw translation per hour. It is available to all Commission officials via the web and internal email. The service is also used by staff in other EU institutions and by public administrations in the Member States. Translations are usually returned to the requester within minutes.

Extensive use has been made of machine translation over the past few years: for example, 860 314 pages were submitted to the system in 2005. The Commission was the main user, with 671 500 pages, some 40% of which (267 182 pages) were requested by the DGT itself. The DGT customises the results by feeding in terminology tailored to its own specific translation needs. Customisation is particularly advanced for English-Spanish and French-Spanish, which are the most popular language pairs among translators, followed by English-French. Translators who use machine translation do so in order to obtain a rough draft translation that they then edit to the quality of a human translation.

In terms of individual requesters, however, the majority of Commission users are administrators in other departments, who most often use the combinations between French, English and German. Administrators principally use machine translation as a multilingual browsing tool, but it also serves as a stop-gap when translations are required at very short notice. In the latter case, the raw machine translation should always be corrected.

For administrators who do not have the time or linguistic ability to correct machine translations themselves, the DGT offers a Rapid Post-Editing Service. This is a network of freelance translators who are experienced in revising raw machine translations. However, as the emphasis is on speed and accuracy rather than on style or terminology, this service can only be used for internal documents. If texts are intended for external distribution, administrators must always ask for a full human translation.

For administrators who do not have the time or linguistic ability to correct machine translations themselves, the DGT offers a Rapid Post-Editing Service. This is a network of freelance translators who are experienced in revising raw machine translations. However, as the emphasis is on speed and accuracy rather than on style or terminology, this service can only be used for internal documents. If texts are intended for external distribution, administrators must always ask for a full human translation.

Source: http://ec.europa.eu/dgs/translation/bookshelf/brochure_en.pdf
(last accessed 6 March 2009 – March 2007 version)



Directorate-General for Translation (DGT) English

European Commission > Translation DG > External Translation

What's new? | Site map | Index | About this site | FAQ | Contact | Search EUROPA

External translation

Contact

- Contact the External Translation Unit

Applying to work with us

- Freelance work
- Calls for tenders and expressions of interest

Tips and guidelines

Language aids for EU translations

Invoicing

- Translations within / outside the EU
- Post-editing work within / outside the EU

Freelance work

In 2007 about 24% of the total output of the Translation DG was done by freelance translators both inside and outside the European Union. This comes to roughly 418 000 pages in 2007, and consists of material which does not have to be translated in-house because it is not confidential, politically sensitive or wanted urgently. The main requirement is for freelancers who can offer the official EU languages, but translation from and into other languages (e.g. Russian) is sometimes needed.

For practical reasons, the DGT **does not recruit individually** but collectively by regularly organising invitations to tender following which framework contracts are concluded with the selected tenderers. The calls for tenders are published in the Official Journal of the European Union and on this website.

Both individual translators and translation agencies can apply. The **languages needed** are primarily the **EU's official languages** (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish) and the **candidate countries' languages** (Turkish and Croatian) mainly in combination with English, French or German. All other languages together make up much less than 1% of the External Translation Unit's workload.

We also issue calls for tenders for services other than translation proper, covering **computer-assisted translation, amendment, revision, editing and post-editing** of output from the Commission's machine translation system.

Very occasionally we award a **one-off contract** for a single piece of work to someone who is registered with us only for a very unusual language combination or a highly specialised subject area.

We also issue calls for tenders for services other than translation proper, covering **computer-assisted translation, amendment, revision, editing and post-editing** of output from the Commission's machine translation system.

Source: http://ec.europa.eu/dgs/translation/workingwithus/freelance/index_en.htm
(last accessed 6 March 2009)



MT post-editing: conclusion 68

“recent activities by localization and translation agencies [...] that use MT systems for translating texts [...] indicate that a market for full post-editing may in fact be underway.”
(Allen, 2003: 306)

“The key issue is how much of the total effort can be handled by a computer and how much must still be done by human labor. Text input, pre-editing, and post-editing can take as much human time and effort as complete human translation.”
(Henisz-Dostert, Macdonald & Zarechnak, 1979: 81)



References and readings (textbooks and online sources) 69

- One chapter from Somers, H. (ed.) (2003) *Computers and Translation: A Translator's Guide*. Amsterdam and Philadelphia, John Benjamins, i.e. + 16 (Jeffrey Allen): "Post-editing", pages 297-317
- Aston, G. (2011) "Tecniche per migliorare la traduzione automatica: post-editing e pre-editing". Bersani Berselli, G. (a cura di) *Usare la traduzione automatica*. Bologna: CLUEB. Capitolo 2 (pp. 33-62)
- Hutchins, W.J. & H.L. Somers (1992) *An Introduction to Machine Translation*. London: Academic Press. Available online at <http://www.hutchinsweb.me.uk/IntroMT-TOC.htm> (various chapters, which can be downloaded, provide further information on the topics discussed in the slides)
- Arnold, D.J., L. Balkan, S. Meijer, R. Lee Humphreys & L. Sadler (1994) *Machine Translation: an Introductory Guide*. London: Blackwells-NCC. Available online at <http://www.essex.ac.uk/linguistics/clmt/MTBook> (various chapters, which can be downloaded, provide further information on the topics discussed in the slides)



References from the slides 70 *(you don't have to read/study these!)*

- Hennisz-Dostert, B., R.R. Macdonald & M. Zarechnak (1979) *Machine Translation*. Mouton Publishers.
- Petrits, A., F. Braun-Chen, J.M. Martínez García, C. Ross, R. Sauer, A. Torquati & A. Reichling (2001) "The Commission's MT System: Today and Tomorrow". In B. Maegaard, B. (ed.) *Proceedings of the MT Summit VIII*. European Association for Machine Translation.
- Senez, D. (1998a) "The Machine Translation Help Desk and the Post-Editing Service". *Terminologie & Traduction*, 1, 1998. European Commission: OPOCE.
- Senez, D. (1998b) "Post-editing service for machine translation users at the European Commission". In *Proceedings of Translating and the Computer 20*. Aslib.
- Wagner, E. (1985) "Post-editing Systran – A challenge for Commission Translators". *Terminologie & Traduction*, 3, 1985. European Commission: OPOCE.



Main scenarios for the use of MT

71

• Information assimilation

- many SLs, only one TL
- unpredictable style
- unpredictable topic / domain
- the MT user is the reader / receiver
- post-editing is possible

• Information dissemination

- only one SL, many TLs
- style can be controlled
- one topic / domain (at a time)
- the MT user is the author / writer
- post-editing is unlikely



Limit input domain / topic

72

- We already know that it is impossible to create MT systems that can offer **fully automatic**, **high-quality** translations for **unrestricted texts**
- So we can try to keep only two out of these three requirements:
 - **total automation of the translation process**
 - **high quality of the output (target text)**
- There are two possibilities to **limit the texts / language** in / for MT:
 - adopt a **controlled language (restricted input)**
 - use the **sublanguage** approach
- Common aims with both options (to the advantage of MT):
 - limited vocabulary (less ambiguity, fewer homographs)
 - more certainty on interpretation (world knowledge and role of context)
 - reduce syntactic variation



Controlled language (1/2)

73

- Prescriptive rules aimed at normalising the style of the input (ST), e.g.
 - do not write sentences with more than 20 words (general, language-neutral)
 - avoid passive constructions, use only active verb forms
 - avoid anaphoras, make all subjects and pronominal references explicit
 - *in EN*: do not omit “that” in relative clauses (language-specific)
 - *in IT*: do not use “solo” as an adverb, but use “soltanto/solamente”
 - *in IT*: use the word “minuto” only as a noun (i.e. to mean 60 seconds);
for the adjectival meaning, use only “piccolo”
 - etc.

The result of controlled language is restricted input



Controlled language (2/2)

74

- Heavily used in technical writing (even without MT)
- It improves the consistency and readability of ST (for humans!)
 - the text is “more precise”, ambiguity is reduced (or removed altogether)
- It simplifies MT into a number of different TLs from the same ST
(which is written in controlled language)
- Authors/writers can find it difficult to apply controlled language rules...
- ... but they can be obtained/adopted/enforced with the help of tailor-made writing aid tools (e.g. *style checkers* ~ Word’s grammar checker)



Sublanguage (1/2)

75

- Natural/normal behaviour of language within a well-defined domain (~ LSP, specialised language, jargon, etc.)
 - “sub-” in the mathematical sense as in “subset”, not derogatory!
 - referred to very well-defined, enclosed, limited domains and texts
 - no need to impose external/explicit rules, language is used that way
- A sublanguage exists and is used regardless of MT, but one can design an MT system that takes advantage of this sublanguage
 - vocabulary
 - limited (relatively few concepts to be covered/expressed)
 - finite/closed (innovation/deviation tend to be avoided)
 - a few homographs, in general limited use of synonyms and coreferences
 - syntax
 - limited range of structures and constructions (regularity + repetitiveness)
 - usually sublanguages are very similar cross-linguistically between SL/TL(s)



Sublanguage (2/2)

76

Sublanguage: “a language used to communicate in a specialized technical domain or for a specialized purpose [...]. Such language is characterised by the high frequency of specialized terminology and often also by a restricted set of grammatical patterns. The interest is that these properties make sublanguage texts easier to translate automatically.”

(Arnold et al., 1994)

- For example, in weather forecast bulletins...
 - predictable vocabulary, precise and unambiguous
 - ditto for syntax (e.g. interrogative structures are spontaneously absent)
 - similar features across languages (IT, EN, DE, FR, ES, RU, etc.)



The Météo MT system (introduction)

77

- Typical example of a sublanguage-based MT system
 - developed from 1965 at Montreal University
 - launched in 1977 to serve the whole of Canada (bilingual country)
 - MT system to translate exclusively weather forecasts (and nothing else!)
 - translations from English (input) into French (output)
 - crucial sector, but not very popular among professional translators (!)
 - input in strictly standard/repetitive format... ideal conditions for MT!
- Sublanguage of weather forecasts (very similar between EN and FR)
 - absence of anaphoric/pronominal references, relative clauses, passives
 - minimum problems of syntactic complexity/ambiguity
 - very specific, unambiguous vocabulary (sensitive to season/conditions)
 - mostly a few verbal moods/tenses used, others very rare
 - telegraphic style: short sentences, parataxis, juxtaposition, ellipsis
 - very similar textual structure and conventions between SL and TL



An example of weather forecast

78

Long Term Forecast Updated: Wednesday, December 10, 2008, 16:56 EST

	Friday Dec. 12	Saturday Dec. 13	Sunday Dec. 14	Monday Dec. 15	Tuesday Dec. 16	Wednesday Dec. 17
	Light snow	Mainly sunny	Cloudy with sunny breaks	Light rain	Scattered showers	Sunny
P.O.P.	90%	0%	20%	90%	90%	0%
High	-2°C	-13°C	-3°C	5°C	0°C	-13°C
Low	-9°C	-15°C	-18°C	-4°C	-1°C	-16°C
Wind	W 10 km/h	SW 10 km/h	SW 10 km/h	SW 20 km/h	W 15 km/h	NW 10 km/h
24-Hr Rain	-	-	-	-	-	-
24-Hr Snow	5-10 cm	-	-	-	-	-

Tendance à long terme Émis le : Mercredi 10 décembre 2008, 15:00 HNE						
	Vendredi 12 décembre	Samedi 13 décembre	Dimanche 14 décembre	Lundi 15 décembre	Mardi 16 décembre	Mercredi 17 décembre
	Faible neige	Généralement ensoleillé	Nuageux avec éclaircies	Faible pluie	Nuageux avec éclaircies et averses dispersées	Ensoleillé
P.D.P.	90%	0%	20%	90%	90%	0%
T. Max	-2°C	-13°C	-3°C	5°C	0°C	-13°C
T. Min	-9°C	-15°C	-18°C	-4°C	-1°C	-16°C
Vents	O 10 km/h	SO 10 km/h	SO 10 km/h	SO 20 km/h	O 15 km/h	NO 10 km/h
Pluie 24 H.	-	-	-	10-15 mm	moins de 1 mm	-



A more discursive example of weather forecast

79

- **Current Weather Forecast**
- **Issued: 11.00 AM GMT Wednesday 10 December 2008**
- **Today**
Day: A mix of sun and cloud. Wind becoming west 20 km/h gusting to 40 early this afternoon. High plus 3.

Night: A few clouds. Wind west 20 km/h gusting to 40 becoming light this evening. Low minus 9.
- **Thursday**
Sunny with cloudy periods. High zero.
- **Friday**
Cloudy with 70 percent chance of flurries. Low minus 10. High minus 1.
- **Saturday**
Periods of snow. Low minus 20. High minus 19.



“Behind the scenes” of Météo

80

- Internal “transfer” (2nd generation) architecture/design
- All the possible morpho-lexical variants included within the system
- 3 separate bilingual dictionaries with correspondences/substitutions, i.e.:
 - general weather-related terms
 - (semi-)fixed sequences of 2 or more words (~collocations, phraseology)
 - geographic names (to be translated, if not included left unchanged in TL)
- Météo’s performance
 - great success rate: 95-100% correct output, mistakes almost invariably due to formal problems in the input, e.g. typos (same standard as HT!)
 - QA and post-editing, when needed, performed by bilingual meteorologists
 - 1989 launch of Météo FR>EN (bulletins of Quebec Weather Office)
 - easy to “migrate”: Météo-96 (for the Atlanta Olympics in the USA)
- (A few) other sublanguage-based MT systems
 - Aviation (aircraft hydraulic systems, EN-FR)
 - TITUS (abstracts from the textile industry, EN/FR/DE/ES)
 - Smart (job offers/ads, EN-FR)
 - TRADEX (military telexes, FR-DE)
 - USSR (patents, EN-RU)



References and readings (textbooks and online sources) 81

- Two chapters from Somers, H. (ed.) (2003) *Computers and Translation: A Translator's Guide*. Amsterdam and Philadelphia, John Benjamins, i.e.
 - + 14 (E. Nyberg et al.): "Controlled language for authoring and transl.", pp. 245-281
 - + 15 (H. Somers): "Sublanguage", pp. 283-295
- Gaspari, F. e E. Zanchetta (2011) "Scrittura controllata per la traduzione automatica". Bersani Berselli, G. (a cura di) *Usare la traduzione automatica*. Bologna: CLUEB. Capitolo 4 (pp. 63-87)
- Hutchins, W.J. & H.L. Somers (1992) *An Introduction to Machine Translation*. London: Academic Press. Available online at <http://www.hutchinsweb.me.uk/IntroMT-TOC.htm> (various chapters, which can be downloaded, provide further information on the topics discussed in the slides)
- Arnold, D.J., L. Balkan, S. Meijer, R. Lee Humphreys & L. Sadler (1994) *Machine Translation: an Introductory Guide*. London: Blackwells-NCC. Available online at <http://www.essex.ac.uk/linguistics/clmt/MTBook> (various chapters, which can be downloaded, provide further information on the topics discussed in the slides)



Computer-assisted translation (CAT) tools: software to help translators

82

- **Computer-assisted translation** or **computer(machine)-aided translation (CAT)** refers to a variety of tools, a family of software products designed to help and support professional translators in their work
- CAT is a "recent" development, derived from MT over the last 20 years
- The idea to develop tools to help and assist (rather than replace) human translators was first proposed by Martin Kay (1980): "PPMMILT" paper
- Later Alan K. Melby and others gave shape to this initial idea
- The actual development of commercial **CAT tools** started in the 1990's – the so-called "**translator's workstation / workbench**", which includes
 - **terminology management packages**
 - **translation memory (TM) software** (+ **text alignment software**, etc.)
- CAT tools are pieces of software designed to enhance the work of translators (maximise productivity and improve quality of translations)
- CAT is relevant only to technical/specialised translation (not literature)



CAT tools, example 1:
terminology management packages

83

- Used to create, store, retrieve and manipulate bi-/multilingual glossaries
- Terminology records typically have a number of flexible fields
- Some are default (pre-defined) fields, others can be customised
- Terminological equivalents, definitions, contextual notes, explanations, grammatical information, references to sources, images, examples, etc.
- Term records are searchable, either manually or automatically
- Terminology management packages are either stand-alone or integrated with other CAT tools, especially with *translation memory software*
- CAT tools sold as *suites* of programmes, e.g. *text alignment software* (matches ST-TT sentences into translation units) and *conversion filters*



CAT tools, example 2:
translation memory (TM) software

84

- **Translation memory (TM):**

“multilingual text archive containing [...] multilingual texts, allowing storage and retrieval of aligned text segments against various search conditions”

(EAGLES* 1995)

* *Evaluation of Natural Language Processing Systems*

- This roughly means: a “filing cabinet” (i.e. a database) of old translations whose bits can be retrieved and used when / as needed by the translator
- Key idea: recycle similar past translations, never translate same text twice
- **Translation memory** indicates both the software tool and the contents of the database, i.e. the whole set of aligned text segments that it includes



Translation memory (TM) software

85

- Essentially a TM is a textual database that can be searched
- Requirements to use TMs
 - source text must be in (or turned into, e.g. with OCR) electronic / digital format
 - translator must have (or get from client) a relevant TM - languages and domain
 - repetitive texts requiring consistency and accuracy, e.g. according to approved phraseology, client's specifications: technical text types (e.g. user manuals), legal documents (e.g. contracts, calls for tenders), etc. + updated material
- Key concepts
 - **translation unit (TU)**: aligned segments of text, pair of sentences in SL & TL
 - **similarity or acceptability threshold**: search parameter set by the translator
 - **translation candidates**: *different types of matches (according to percentage of similarity between sentence to be translated and segment of TU in source lang)*



Example of aligned bilingual translation unit (English/Italian) in XML format

86

```
1 <tu creationdate="20080424T135938Z" changedate="20080514T172828Z" creationid="Utente SSLMIT 12" changeid="Utente
  SSLMIT 16" srclang="en-US">
2   <tuv xml:lang="en-US">
3     <seg>She is the author of numerous articles regarding learning disabilities and she speaks often before
      parent and teacher groups concerning learning and behavior problems.
4     </seg>
5   </tuv>
6   <tuv xml:lang="it-IT">
7     <seg>È autrice di numerosi articoli riguardanti le disabilità di apprendimento e ha tenuto spesso
      conferenze davanti a gruppi di genitori e insegnanti sui problemi del comportamento e dell'apprendimento.
8     </seg>
9   </tuv>
10 </tu>
11 <tu creationdate="20080428T165122Z" changedate="20080428T165122Z" creationid="Utente SSLMIT 12" changeid="Utente
  SSLMIT 12" srclang="en-US">
12   <tuv xml:lang="en-US">
13     <seg>All expectations need to be direct and explicit. Don't require this child to "read between
      the lines" to glean your intentions.
14     </seg>
15   </tuv>
16   <tuv xml:lang="it-IT">
17     <seg>Esplicitare chiaramente tutte le aspettative, in modo da non richiedere al bambino di "leggere tra le
      righe" per cogliere le intenzioni.
18     </seg>
19   </tuv>
20 </tu>
```



Using translation memory (TM) software 1/4

87

- Scenario
 - translate the user manual of a printer (new model) from English into Italian
 - a lot of repetition within each document (manual) – see examples on p. 239
 - overlap and repetitions across updated (old-new) versions of the documentation
 - you have translated the previous manual(s), so you have the relevant TM
- Process
 - activate the relevant TM for lang. pair (similar topic / domain / texts / clients)
 - set level of matching required for past units to be shown (say, 75% similarity)
 - while you translate the TM retrieves and shows you units above 75% similarity
 - TM can run in parallel with integrated terminology management packages (which suggest good TL term equivalents for specific words while translating) as well as machine translation software for first-draft translation
- The translator is always in charge, he/she decides which translation candidates (matches) to use, either as they are or modifying them
- The TM software is there to help the translator, who remains central



Using translation memory (TM) software 2/4

88

- Translation of a printer manual English (A) → Italian (B)

Source text (in language A)

ST: There are ④ ways to change print settings for this printer

Exact match (everything in the segment is exactly the same)

A: There are 4 ways to change print settings for this printer

B: Ci sono 4 modi per cambiare le impostazioni di stampa di questa stampante

Full match (only figures, dates and similar small details are different)

A: There are ② ways to change print settings for this printer

B: Ci sono ② modi per cambiare le impostazioni di stampa di questa stampante



- New feature, recently added to (some) TM tools

Source text (in language A)

ST0: How to change default printer settings

ST1: There are 4 ways to change print settings for this printer

This is the sentence that you're currently trying to translate in your ST (above, ST1) and stored in the TM (below, A1)

This is the segment coming before the sentence to be translated both in the ST and in the TU, which provides context

Context match (everything is the same, including previous segment)

A0: How to change default printer settings

A1: There are 4 ways to change print settings for this printer

B1: Ci sono 4 modi per cambiare le impostazioni di stampa di questa stampante



Source text (in language A)

ST: "There are 4 ways to change print settings for this printer"

Fuzzy match 85% similar (a few words in translation unit are different)

A: "There are **several** ways to change print settings for **the** printer"

B: "Ci sono **vari** modi per cambiare le impostazioni di stampa **alla** stampante"

Fuzzy match 60% similar (some words in translation unit are different)

A: "There are **several** ways to **modify** the default setting **of your** printer"

B: "Ci sono **vari** modi per **modificare** l'impostazione **standard** della tua stampante"

- With the acceptability threshold of the TM tool set at 75%, no candidate translation unit under that level of similarity is retrieved and shown to the translator!!



TM leverage: compare current ST against relevant TM database (to estimate time and give quote) 91

- Material that can be “recycled”, i.e. available translation units that can be potentially reused because they are to some extent similar to (parts of) the new text to be translated, grouped into levels of match similarity
- Example of ***leverage*** analysis (done before translating):

Match types	Segments	# words	%
<i>Repetitions</i>	141	1.329	8
<i>100% (Exact Match)</i>	167	635	4
<i>95-99%</i>	22	246	1
<i>85-94%</i>	36	379	2
<i>75-84%</i>	66	621	4
<i>50-74%</i>	27	290	1
<i>0% (No Match)</i>	869	11.740	80
<i>TOTAL</i>	1.328	15.240	100

- ***Pre-translate*** function: inserts (partial) matches into target document



Main sources

92

- Three chapters from Austerlühl, F. (2001) *Electronic Tools for Translators*. Manchester: St. Jerome Publishing, i.e.
 - + 1 “Translation in the information age”, pages 1-17
 - + 7 “Computer-assisted terminology management”, pages 102-123
 - + 9 “Déjà Vu? – Translation memories and localization tools”, pages 134-152
- Four chapters from Bowker, L. (2002) *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press, i.e.
 - + 1 “Why do translators need to learn about translation technology?”, pages 11-21
 - + 4 “Terminology-Management Systems”, pages 77-91
 - + 5 “Translation-Memory Systems”, pages 92-127
 - + 6 “Other New Technologies and Emerging Trends”, pages 129-140



Main sources (continued)

93

- Three chapters from Somers, H. (ed.) (2003) *Computers and Translation: A Translator's Guide*. Amsterdam and Philadelphia: John Benjamins, i.e.
 - + 2 (H. Somers): "The translator's workstation", pages 13-30
 - + 3 (H. Somers): "Translation memory systems", pages 31-47
 - + 4 (L. Bowker): "Terminology tools for translators", pages 49-65
- Five chapters from Quah, C.K. (2006) *Translation and Technology*. Basingstoke, Palgrave MacMillan, i.e.
 - + 1 "Definition of Terms", pp. 6-21
 - + 4 "Computer-Aided Translation Tools and Resources", pp. 93-128
 - + 5 "Evaluating Translation Tools", pp. 129-151
 - + 6 "Recent Developments and Future Directions", pp. 152-171
 - + 7 "Translation Types Revisited", pp. 172-196
- Lecci, C. & E. Di Bello (2012) *Usare la traduzione assistita*. Bologna: CLUEB



Computer-assisted translation (CAT) tools – online resources on translation memory software

94

- MeLLANGE (Multilingual eLearning in Language Engineering) was a European research project (ended in 2007) that brought together academic and industry partners from France, Austria, the Czech Republic, Germany, Italy, Spain, Switzerland and the UK to create innovative e-learning materials for trainee and professional translators
- The project consortium designed and implemented a variety of interactive courses, focusing for example on the use of translation memory software and corpora for translators
- Log on to this webpage which gives access to the e-learning materials developed by the MeLLANGE project (divided into topics and units):
 - <http://parles.upf.edu/llocs/moodle/course/category.php?id=122>
- This is a gateway to the MeLLANGE resources, from which you can choose which topics you would like to investigate



***Computer-assisted translation (CAT) tools –
online resources on translation memory software***

95

- Click on the link “Translation Memory” at the bottom
- You can consult the materials as a guest, without logging in as a user
- This e-learning course on (the use of) translation memory software is broken down into 11 self-contained units (or lessons) which can be completed in a step-by-step fashion and contain glossaries, explanations, tutorial-style animations, tests, quizzes, exercises, etc.
- Use these e-learning materials on translation memory software to consolidate what was covered in class, and to familiarise yourself with the English terminology that is required to discuss CAT tools (for the exam!)



***Self-access to state-of-the-art translation memory tools
SDL Trados Studio 2011 and memoQ by Kilgray***

96

- MA students have access to SDL Trados Studio 2011 and memoQ on the computers of Lab 1.6, on the first floor of Palazzo Ciccolini
- So you can try for yourself and learn to use state-of-the-art TM software
- For SDL Trados Studio 2011, see the book by Lecci & Di Bello (2012)
- memoQ is an integrated translation environment; more info at:
 - <http://kilgray.com/products/memoq>
- Profile of Kilgray, the Hungary-based company producing memoQ:
 - <http://kilgray.com/company>
- To learn more about memoQ, the Resource Center section of Kilgray’s website has online training videos, user guides, documentation and further information in English which will help you to understand the functionalities of this tool and will show you how to use it:
 - <http://kilgray.com/resource-center>



OPTIONAL SEMINAR
***Linguistic and professional issues
in the use of TM software***

97

- Features of texts for which the use of CAT tools and particularly TMs is especially appropriate and beneficial
 - highly/clearly structured, well formatted
 - repetitive (within same document and/or across updates/new versions)
 - accuracy and consistency in TL renderings are required
 - concrete examples of relevant text types, genres and documents
 - technical/specialised texts such as manuals, contracts, financial reports, etc.



***Linguistic and professional issues
in the use of TM software***

98

- You can build TM databases in two ways
 - incrementally: start with empty database, and feed TUs into it as you go along
 - utilising old translations (including ones on paper) with text alignment software
- TMs maximise productivity (via reuse) & improve quality (consistency)
- Freelance (technical / specialised) translators expected to use CAT tools
- Translation agencies hire translators who can use CAT software
- It is a way to offer better services and keep clients returning in the future
- Risky for freelancers to reject CAT technology, especially in some fields
- Clients (e.g. translation agencies) expect translators to use CAT tools
- Who owns and can use the TM databases? (*confidentiality agreements*)
- Fair that translators are paid different rates for exact/full/fuzzy matches?



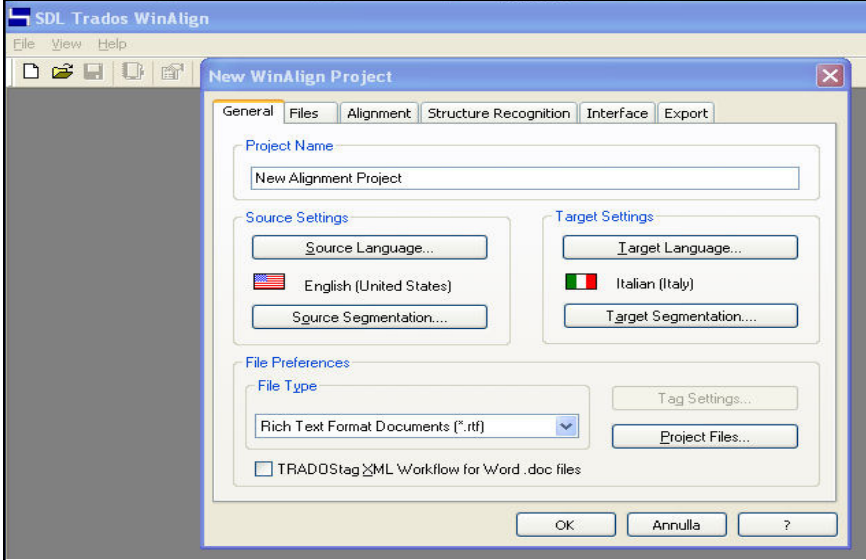
Linguistic issues in the use of TM software

99

- Alignment procedure – an example (using SDL Trados WinAlign)
- We simulate the alignment of an old translation (Eng-Ita)
- A company's annual financial report (Ente Fiera di Milano)
- Original in Italian, to be translated into English (lang. direction irrelevant)
- We have already translated the 2007 report (but without using TM!)
- We get the commission for the new 2008 edition of the report and we want to translate it recycling our past work, by using TM software this time, so we need to populate a TM database (documents on pp. 240-241)
 - highly/clearly structured (makes alignment easier, maximises leverage)
 - repetitive (both internally within document and across different years)
 - important to adopt accurate and consistent renderings (style, phraseology)

A	EN: Warning: FR: Avertissement : <i>Titles = 1:1</i> 100
B	EN: This computer program is protected by copyright law and international treaties. FR: Ce logiciel est protégé par les lois et les traités internationaux sur le droit d'auteur. <i>Sentences = 1:1</i>
C	EN: Unauthorized reproduction or distribution of this program, or any portion of it, may result in severe civil and criminal penalties, and will be prosecuted to the maximum extent possible under the law. FR: Toute reproduction ou distribution partielle ou totale, par quelque moyen que ce soit, est strictement interdite. Toute personne ne respectant pas ces dispositions se rendra coupable du délit de contrefaçon et sera passible des sanctions pénales prévues par la loi. <i>Sentences = 1:2</i>
D	EN: The "0" button and the "1" option affect the current application. The other options affect all applications. FR: Les modifications apportées par le bouton « 0 » et l'option « 1 » n'affecteront que l'application en cours alors que les autres options seront répercutées dans toutes les applications. <i>Sentences = 2:1</i>

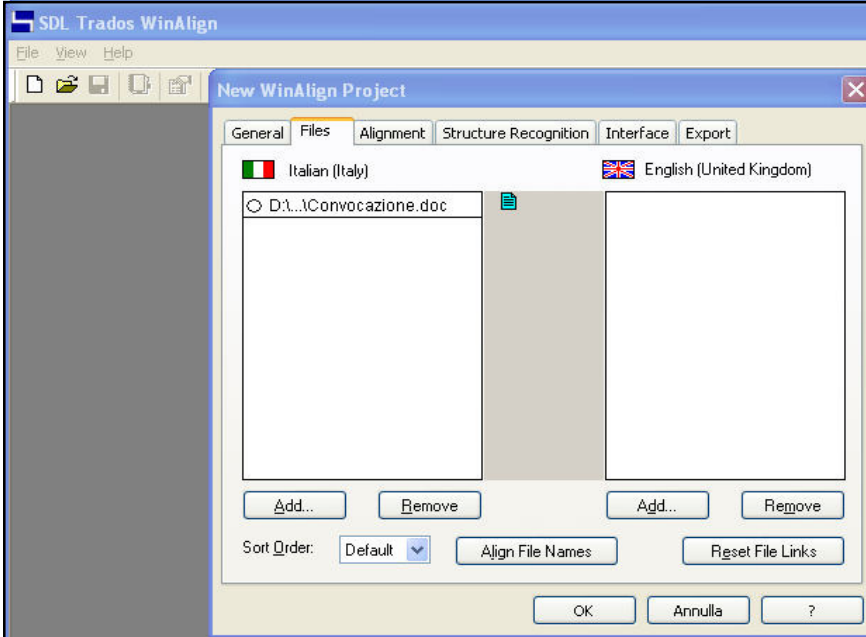
101



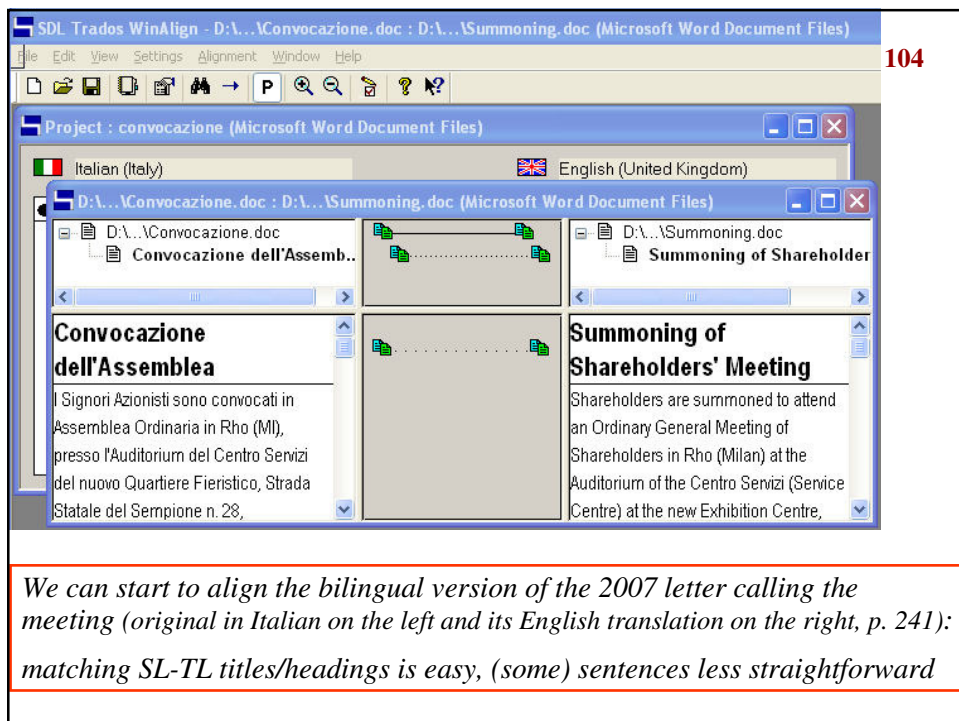
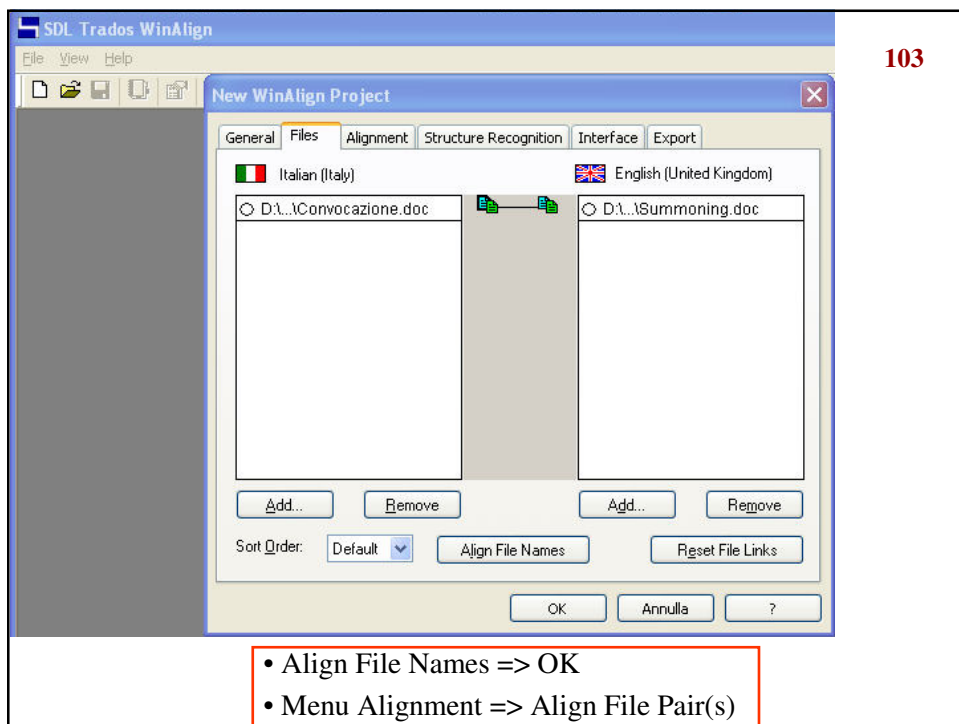
- Select languages involved (SL and TL)
- Select format/type of files

CAT tools are equipped with filters to handle/convert a variety of file formats

102



- Upload/select the two files with the texts that you wish to align



Convocazione dell'Assemblea

Signori Azionisti sono convocati in Assemblea Ordinaria in Rho (MI), presso l'Auditorium del Centro Servizi del nuovo Quartiere Istituzionale, Strada Statale del Sempione n. 28, (disponibilità di parcheggio riservato con ingresso da Porta Sud), per il giorno 27 aprile 2007 alle ore 10:00 in prima convocazione e, occorrendo, in seconda convocazione, per il giorno 28 aprile 2007, stessi ora e luogo, per deliberare sul seguente

Ordine del giorno

1. Bilancio al 31 dicembre 2006; Relazione del Consiglio di Amministrazione sulla gestione e Relazione del Collegio Sindacale; deliberazioni inerenti e conseguenti.

2. proroga della durata dell'incarico di revisione alla società PricewaterhouseCoopers SpA, in adeguamento al novellato art. 58, comma 4, del D.Lgs. 24.02.1998, n. 58 e conseguente allineamento del relativo compenso; deliberazioni inerenti e conseguenti.

Summoning of Shareholders' Meeting

Shareholders are summoned to attend an Ordinary General Meeting of Shareholders in Rho (Milan) at the Auditorium of the Centro Servizi (Service Centre) at the new Exhibition Centre, Strada Statale del Sempione 28, (reserved parking space available with entry from Porta Sud - the South Gate) at 10 a.m. on April 27th 2007 on first call and, if a second call is necessary, on April 28th 2007 at the same time and place to discuss and vote on the following

Agenda

1. Year-end financial statements as at December 31st 2006, Directors' Management Report, and Report of Statutory Auditors' Committee; relevant and consequent resolutions.

2. Extension of the term of office of the auditing agreement with PricewaterhouseCoopers SpA, in line with the new Article 159 paragraph four of Legislative Decree no. 58 of February 24th 1998, and resulting realignment of the relative fee; related and consequent decisions.

105

100%	Committed alignment units	_____
80 – 99%	Reliable alignment units	
60 – 79%	Probable alignment units	
1 – 59%	Less probable alignment units	

Levels of confidence in the correctness of the automatic alignment for specific parts of text, based on statistical estimation of accuracy

Assemblea

in Assemblea Ordinaria in Rho (MI), presso l'Auditorium del Centro Servizi del nuovo Quartiere Istituzionale, Strada Statale del Sempione n. 28, (disponibilità di parcheggio riservato con ingresso da Porta Sud), per il giorno 27 aprile 2007 alle ore 10:00 in prima convocazione e, occorrendo, in seconda convocazione, per il giorno 28 aprile 2007, stessi ora e luogo, per deliberare sul seguente

Ordine del giorno

1. Bilancio al 31 dicembre 2006; Relazione del Consiglio di Amministrazione sulla gestione e Relazione del Collegio Sindacale; deliberazioni inerenti e conseguenti.

2. proroga della durata dell'incarico di revisione alla società PricewaterhouseCoopers SpA, in adeguamento al novellato art. 58, comma 4, del D.Lgs. 24.02.1998, n. 58 e conseguente allineamento del relativo compenso; deliberazioni inerenti e conseguenti.

Summoning of Shareholders' Meeting

Shareholders are summoned to attend an Ordinary General Meeting of Shareholders in Rho (Milan) at the Auditorium of the Centro Servizi (Service Centre) at the new Exhibition Centre, Strada Statale del Sempione 28, (reserved parking space available with entry from Porta Sud - the South Gate) at 10 a.m. on April 27th 2007 on first call and, if a second call is necessary, on April 28th 2007 at the same time and place to discuss and vote on the following

Agenda

1. Year-end financial statements as at December 31st 2006, Directors' Management Report, and Report of Statutory Auditors' Committee; relevant and consequent resolutions.

2. Extension of the term of office of the auditing agreement with PricewaterhouseCoopers SpA, in line with the new Article 159 paragraph four of Legislative Decree no. 58 of February 24th 1998, and resulting realignment of the relative fee; related and consequent decisions.

106

Edit Segment

Join Segments

Split Segment

Insert Segment

Start Connecting Segments

Finish Connecting Segments

Disconnect

Commit

Realign Structure Level from Active Unit

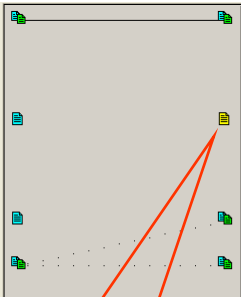
Selected Units

All Non-Committed Units

All Units

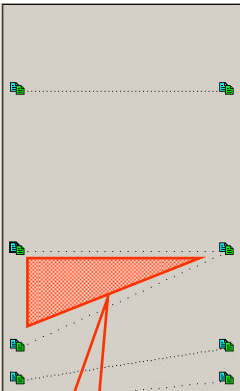
- Edit segment => Quick edit (/ Advanced edit)
- Insert Segment
- Disconnect, Commit
- Realign...


107

Convocazione dell'Assemblea I Signori Azionisti sono convocati in Assemblea Ordinaria in Rho (MI), presso l'Auditorium del Centro Servizi del nuovo Quartiere Fieristico, Strada Statale del Sempione n. 28, (disponibilità di parcheggio riservato con ingresso da Porta Sud), per il giorno 27 aprile 2007 alle ore 10.00 in prima convocazione e, occorrendo la seconda convocazione, per il giorno 28 aprile 2007, stessi ora e luogo, per deliberare sul seguente Ordine del giorno Bilancio al 31 dicembre 2006, Relazione del Consiglio di Amministrazione sulla gestione e Relazione del Collegio sindacale; deliberazioni inerenti e conseguenti.		Summoning of Shareholders' Meeting Shareholders are summoned to attend an Ordinary General Meeting of Shareholders in Rho (Milan) at the Auditorium of the Centro Servizi (Service Centre) at the new Exhibition Centre, Strada Statale del Sempione 28, (reserved parking space available with entry from Porta Sud - the South Gate) at 10 a.m. on April 27th 2007 on first call and, if a second call is necessary, on April 28th 2007 at the same time and place to discuss and vote on the following Agenda Year-end financial statements as at December 31st 2006, Directors' Management Report, and Report of Statutory Auditors' Committee; relevant and consequent resolutions.
--	---	--

*“Orphan” segment
in cases of
0:1 or 1:0 alignments*


108

hanno diritto di intervenire all'Assemblea, ai sensi della legge e dello Statuto sociale, gli Azionisti in relazione ai cui possessi azionari sia pervenuta presso la sede legale della Società, o, preferibilmente, presso la sede operativa ed amministrativa in Rho (MI), SS del Sempione, 28, Centro Servizi, reception uffici, almeno due giorni non festivi precedenti la data fissata per la riunione assembleare la prescritta comunicazione rilasciata dal rispettivo intermediario incaricato. La documentazione relativa agli argomenti all'ordine del giorno prevista dalla normativa vigente verrà depositata presso la sede legale della Società e presso la sede operativa ed amministrativa in Rho (MI), SS del Sempione, 28, Centro Servizi, reception uffici, presso Borsa Italiana, a disposizione dei Soci e del pubblico, nei termini prescritti. Gli Azionisti hanno facoltà di ottenere copia di tutta la predetta documentazione. Rho, 26 marzo 2007		In accordance with applicable law and the company's by-laws, the meeting may be attended by all shareholders who have provided the required communication, issued by their respective financial intermediaries, in relation to their respective shares to the company's registered office or, preferably, to the operations and administration offices in Rho, MI (Italy), SS del Sempione no. 28, Centro Servizi, reception offices, at least two business days prior to the date set for said meeting. The documentation concerning the items on the agenda as required by prevailing laws and regulations will be available to shareholders and to the public at the company's registered office, the operations and administration offices in Rho, MI (Italy), SS del Sempione no. 28, Centro Servizi, reception offices, and the offices of Borsa Italiana Shareholders have the right to obtain a copy of all such documentation. Rho, 26 March 2007 The Chairman of the Board of Directors
--	---	---

2:1 alignment

109

The screenshot shows the SDL Trados WinAlign application window. The 'Alignment' menu is open, displaying options: 'Start Connecting Segments', 'Finish Connecting Segments', 'Disconnect', 'Commit', 'Align Structure Level', 'Realign Structure Level from Active Unit', 'Mark as Finished', and 'File Pair Statistics...'. The 'Mark as Finished' option is highlighted. The main window displays two text segments being aligned. The left segment is in Italian, and the right segment is in English. The bottom of the window shows the project statistics.

- Mouse right-click => Commit all Units
- Alignment Menu => Mark as Finished
- Alignment Menu => File Pair Statistics

110

The screenshot shows the 'Project Statistics' dialog box. It contains a table of statistics and buttons for 'Save...', 'OK', 'Cancel', and 'Help'.

Statistics Information:	
Date:	November 06, 2008
Time:	15:48:53
Source File Name:	D:\Personale\corsi_08_09\trad_assist\lez8\win_aligned\Convocazione.doc
Target File Name:	D:\Personale\corsi_08_09\trad_assist\lez8\win_aligned\Summoning.doc
Number of Source Segments:	12
Number of Target Segments:	11
Number of Aligned Units:	11
Number of Committed Units:	11
Number of High Quality Units:	0
Number of Medium Quality Units:	0
Number of Low Quality Units:	0
Number of Unconnected Source Segments:	0
Number of Unconnected Target Segments:	0
Number of 1:2 and 2:1 Units:	1

Summary of alignment task, after the user has finally confirmed ("commit") the alignment done automatically by the text alignment software (SDL Trados WinAlign, in this case) and subsequently manually checked and revised it



111

Edit WinAlign Project

General Files Alignment Structure Recognition Interface **Export**

Format:

☐ Translator's Workbench Import Format

☒ Translation Memory Exchange Format (TMX)

Options:

Export Threshold: 1%

☐ Export as Table ☒ Export as TMX 1.4

☐ Export as Plain Text

☒ Include Creation Date

Creation User:

Save / Export format specific to this TM tool (SDL Trados)

TMX: format to save and export newly created TM database to other TM tools and use it with them

• Settings Menu => Project => Export



112

```
<?xml version="1.0" ?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
<header
  creationtool="TRADOS WinAlign"
  creationtoolversion="7.0"
  segtype="sentence"
  o-tmf="TW4Win 2.0 Format"
  adminlang="EN-US"
  srclang="IT-IT"
  datatype="rtf"
  creationdate="20081106T155819Z"
</header>
<body>
<tu creationdate="20081106T155819Z" creationid="ALIGN!">
<tuv xml:lang="IT-IT">
<seg>Convocazione dell'Assemblea</seg>
</tuv>
<tuv xml:lang="EN-GB">
<seg>Summoning of Shareholders' Meeting</seg>
</tuv>
</tu>
```

TMX (Translation Memory Exchange) format is compatible across different (brands of) TM tools



Linguistic issues in the use of TM software

113

- Types of matches – now you can use your TM database of old aligned SL/TL segments from 2007 to translate the new 2008 annual report!
- *Exact match* (these examples are taken from other documents!):

New source segment	Empty the pipe.
Stored TM unit	EN: Empty the pipe. FR: Videz la pipe.

- *Full match* is similar (only figures, numbers, dates, etc. change)
- *Context match* (preceding segment in SL is the same, too – cf. slide 89)



Multiple fuzzy matches retrieved and ranked (from most to least similar):

114

New source segment	The operation was interrupted because the file was hidden.
Stored TM unit (best match) 85%	EN: The operation was interrupted because the Ctrl-c key was pressed. FR: L'opération a été interrompue car la touche Ctrl-c a été enfoncée.
Stored TM unit (second-best match) 70%	EN: The specified method failed because the file is hidden. FR: La méthode spécifiée a échoué car le fichier est masqué.
Stored TM unit (third-best match) 65%	EN: The operation was interrupted by the application. FR: L'opération a été interrompue par l'application.
Stored TM unit (fourth-best match) 42%	EN: The requested operation cannot be completed because the disk is full. FR: Le disque est saturé. Impossible de continuer l'opération requise.



115

Sub-segment *fuzzy matches* (these are shown only if the similarity / acceptability threshold is set sufficiently low by the translator):

New source segment	First, <u>check for disk space on the drive</u> that contains the Temp folder;
Stored TM unit	EN: 3. Close other programs, <u>check for disk space on the drive</u> you are saving to, and then save again. FR: 3. Fermez des applications, vérifiez l'espace disque disponible sur le lecteur où vous voulez enregistrer votre fichier, puis réenregistrez votre travail.



116

Multiple sub-segment *fuzzy matches* (translation candidates) can be combined to exploit (partial) useful suggestions, but this depends on the similarity/acceptability threshold set by the translator:

New source segment	The file operation <u>cannot be completed because the disk is full.</u>
Sub-segment match	EN: There is not enough memory to <u>perform the file operation.</u> FR: Mémoire insuffisante pour continuer cette opération de fichier.
Sub-segment match	EN: This action <u>cannot be completed because the</u> program is busy. FR: Impossible de continuer cette action car le program est occupé.
Sub-segment match	EN: <u>Disk is full.</u> FR: Le disque est saturé.



Technical and practical matters

117

- Exact / full / context / fuzzy matching takes place at whole sentence level
 - no sub-sentential matches (i.e. more refined/specific granularity)
- What similarity/acceptability threshold should the translator set?
 - trade-off: irrelevant/distracting noise/entropy vs. good/useful candidates
 - need to work with “relevant” TM databases which contain useful matches
- Criteria used to define sentence resemblance and estimate % of matches
 - little sensitivity to actual semantic similarity of sentence/segment/TU content
 - not based on similar meaning of the sentence and its constituents, but...
 - ... calculated on the basis of “difference” between letter/character strings

ST: I bambini giocano in gruppo con il pallone

FM1: I tombini gioovano il grullo con il tallone (90% similar)

FM2: I bimbi stanno giocando a calcio insieme (42% similar)



Linguistic issues in the use of TM software

118

- Issue of SL-TL direction of the segments/TUs stored in memory
 - direction of the SL-TL combination could be largely irrelevant (perhaps...)
- Does using TM tools help translating into the L2 (e.g. English for us)?
 - need to be sure of the good quality of TM database contents (stored TUs)
- Different philosophies and approaches in TM tools on:
 - need to show formatting and layout of text being translated or plain text?
 - formatting and layout are usually meaningful to translators (e.g. sentence or item found in bulleted list vs. whole “proper” sentence), but may distract certain translators sometimes (idea is to focus on text, not form/appearance)
 - “integrated” (usual Word + extra buttons) or proprietary text editor?
 - different ad-hoc text editor could be more powerful, but the translator needs to get used to working in a whole new environment, losing confidence



Professional issues in the use of TM software

119

- Advantages of using TM software
 - for what practical purposes can TM leverage be used?
- Example of *leverage analysis* (done before translating):

Match types	Segments	# words	%
Repetitions	141	1.329	8
100% (Exact Match)	167	635	4
95-99%	22	246	1
85-94%	36	379	2
75-84%	66	621	4
50-74%	27	290	1
0% (No Match)	869	11.740	80
TOTAL	1.328	15.240	100



Professional issues in the use of TM software

120

- Advantages of using TM software
 - *For translators*
 - increased speed (shorter turnaround/delivery time)
 - enhanced quality, e.g. accuracy and consistency (when required)
 - reuse of past work and investment for future similar translations
 - maximised productivity (larger volume of translation per unit of time)
 - can offer more competitive rates (if reused/leveraged matches are discounted)
 - client “fidelisation”
 - *For agencies*
 - same as above + possibility to pay less money for freelancers
 - *For end clients*
 - same as above + possibility to pay less money for translations



Professional issues in the use of TM software

121

- Ownership, use and distribution of TM databases
 - Complex issues of copyright and intellectual property rights
 - complicated even for pen-and-paper translations (different national laws)
 - original (SL) author / translator / agency / publisher / copyright holder
 - Straightforward if the translator is working directly for the end client
 - the translator owns, uses and has every right on the TM database
 - Very difficult if the translator works for an agency or company
 - TM databases can be easily monetised, they have direct commercial value
 - the translator creates/augments the TM, but who owns the rights to exploit it?
 - contracts with confidentiality clauses and provisions for limited use
 - translators normally not entitled to benefit from financial returns of TM use



Professional issues in the use of TM software

122

- “Ethical” problems
 - Actual rates of payment for freelance translators are based on matches
 - e.g. only new material or low matches (e.g. <50% similarity) are paid in full
 - is it fair not to pay at all exact/full matches (translators still check them)?
 - is it fair to pay a reduced amount for high (i.e. very similar) fuzzy matches?
 - TM databases are by-products of the translations, so shouldn’t the (freelance) translators producing them be entitled to some of the profits?
 - The value of TM databases consists in allowing future (similar) translations to be done better, faster (and at a lower cost...)
 - So the translator is caught between a rock and a hard place, as he/she
 - is expected to use TMs to get a job in the first place (forced by agencies)
 - has to invest money upfront to buy his/her own TM/CAT software
 - gets paid less for an assignment if there is substantial leverage
 - doesn’t share the revenues and increased profits of using TM software
 - has increasingly less negotiating power with agencies and end clients



Professional issues in the use of TM software

123

- “Ethical” problems – some food for thought
 - Notion of “collective translator”: when I use TM software I translate by “assembling” (i.e. re-combining) chunks of translations already done previously, possibly by other people... do we like this?!?
 - Some say that translating with TMs limits/kills translator’s creativity
 - Do translators translate differently as a result of using TMs?
 - to what extent are translators influenced by the candidate (fuzzy) matches?
 - forced to translate individual sentences in sequence and separately
 - difficult to be aware of the whole text, one loses sight of (even immediate) context (e.g. anaphoras, inter-sentential co-references, parallelisms, etc.)
 - (un)conscious tendency to fragment translations so they are TM-friendly?



Professional issues in the use of TM software

124

- “Ethical” problems – a summary
 - TMs maximise productivity (via reuse) and improve quality (consistency)
 - Clients (and translation agencies) expect translators to use TM software
 - Translation agencies hire translators who use CAT software
 - some agencies even demand that translators use a specific brand of TM (but TMX format should allow for the compatible exchange of TM databases...)
 - So (freelance) translators must invest in CAT/TM software
 - very risky for freelancers to reject CAT technology, especially in some fields
 - it is a way to offer better services and keep clients returning in the future
 - it is essential for specialised translators who target the high end of the market
 - But there are some serious problems
 - who owns (and has the right to use/exploit) the TM databases?
 - differentiated pricing policies for exact/full/context and high fuzzy matches



Professional issues in the use of TM software

125

- “Ethical” problems – from the point of view of students
 - Teaching/learning the main concepts of how CAT tools and TM software work in general, rather than exposing you to specific brands in class
 - For me it is important that you are aware of these issues/problems and you make your own decisions as to how to position yourselves on the market
 - Have the ability to appreciate the advantages and drawbacks of CAT tools, and know the basic notions to start/learn to use them professionally
 - For example, be able to judge that for certain (types of) texts it can be very beneficial to align past translations and work on them with TM software
 - regard creating TM databases as an investment for the future
 - but avoid having unfair working conditions imposed on you (e.g. by agencies)



Professional issues in the use of TM software

126

- New working scenarios
 - More/unprecedented possibilities to work effectively in team
 - Groups of translators work on (parts of) long documents at the same time
 - share (and contribute to) the same TMs in real time over a network
 - consistency and accuracy are safeguarded and improved as you go along
 - For example: 6 translators work on a chapter each of a financial report
 - same initial TM database and terminology resources made available to them
 - while translators work, updates (added terms, new TUs) are shared by all
 - Practical benefits of this working set-up
 - This can be extended to several language combinations at the same time (e.g. a long financial report in English with 6 chapters has to be translated quickly and efficiently into N languages, by N teams of translators)



Professional issues in the use of TM software

127

- New professional profiles
 - Shift from the point of view of (single/isolated) translation job/assignment to an emphasis on (multiple/connected) *translation projects*
 - re-use relevant assets and recycle resources that are already available
 - invest in the development of new term records and TUs for TM databases to be deployed in the future for similar translation projects
 - Role of the (*translation*) *project manager*, esp. for multilingual projects
 - negotiates project delivery and costs with the client + liaises with them
 - finds and allocates the appropriate resources (TMs), and coordinates their use
 - estimates time and number of professionals required to complete the project
 - monitors the progress of each target language and keeps them synchronised
 - organises quality control and runs quality assurance procedures (e.g. DTP, interim milestones, revisions by senior colleagues, final proof-reading, etc.)
 - takes care of all the resources (e.g. TM database maintenance, updating, etc.)



What is a corpus?

Some (*authoritative*) definitions

128

- “a collection of naturally-occurring language text, chosen to characterize a state or variety of a language” (Sinclair, 1991:171)
- “a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis” (Francis, 1992:7)
- “a closed set of texts in machine-readable form established for general or specific purposes by previously defined criteria” (Engwall, 1992:167)
- “a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration” (McEnery & Wilson, 1996:23)
- “a collection of (1) machine-readable (2) authentic texts [...] which is (3) sampled to be (4) representative of a particular language or language variety” (McEnery et al., 2006:5)



What types of corpora exist?

129

A brief overview

- A **corpus** is a *principled collection of naturally occurring electronic texts designed to be a representative sample of language in actual use*
- Some of the main features and key criteria used to describe and classify corpora:

<i>general</i>	<i>closed / finite</i>
<i>specialised</i>	<i>open-ended (monitor)</i>
<i>written</i>	<i>raw (pre-corpus)</i>
<i>spoken (transcribed)</i>	<i>marked-up (augmented)</i>
<i>balanced</i>	<i>POS-tagged (augmented)</i>
<i>mixed</i>	<i>annotated (augmented)</i>
<i>synchronic</i>	<i>monolingual (see p. 242)</i>
<i>diachronic</i>	<i>bi- / multilingual</i>
<i>static</i>	<i>parallel (see p. 243)</i>
<i>dynamic</i>	<i>comparable</i>



Key concepts and technical notions in corpus-based translation studies

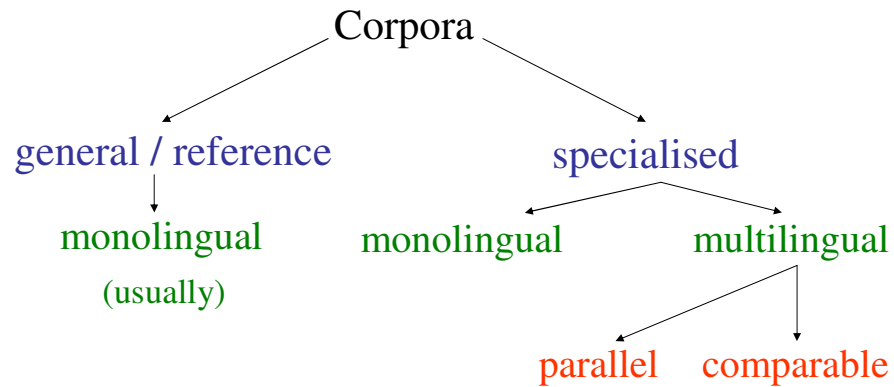
130

- Wordlist, frequency list
- Function/grammatical words vs. content/lexical words (lexical density)
- Types, tokens, type/token ratio (lexical variation)
- Lemmatisation (morphological analysis)
- (POS-)Tagging (grammatical analysis)
- Parsing (syntactic analysis)
- KWIC (keyword in context)
- Concordance (concordancing software) – see pp. 242-243
- Nodeword
- Collocation (collocates)

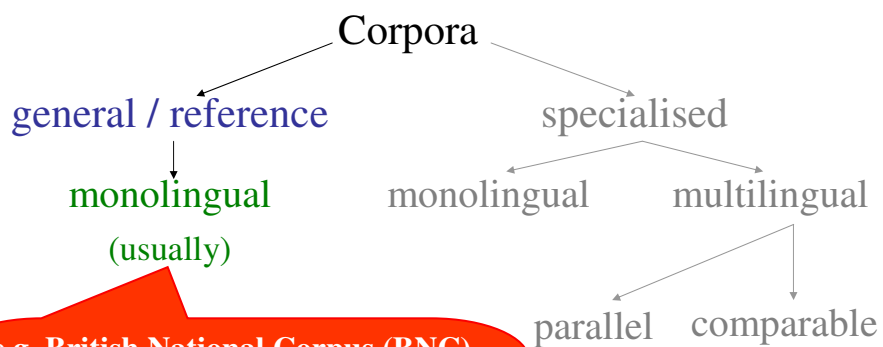


A translation-relevant corpus typology

131



General / reference monolingual corpora (of English) 132



e.g. British National Corpus (BNC)
COCA (AmEng)
Repubblica
Leeds Internet corpora
Mannheim corpora
etc., you normally find them ready

Example:
BNC concordances
for nodewords “eyes”
and “eye” (see p. 242)



Editorial

Nature **455**, 835-836 (16 October 2008) | doi:10.1038/455835b; Published online 15 October 2008

Cut-throat savings

In an attempt to boost its struggling economy, Italy's government is focusing on easy, but unwise, targets. ▲ Top

It is a dark and angry time for scientists in Italy, faced as they are with a government acting out its own peculiar cost-cutting philosophy. Last week, tens of thousands of researchers took to the streets to register their opposition to a proposed bill designed to control civil-service spending (see [page 840](#)). If passed, as expected, the bill would dispose of nearly 2,000 temporary research staff, who are the backbone of the country's grossly understaffed research institutions — and about half of whom had already been selected for permanent jobs.

<http://www.nature.com/nature/journal/v455/n7215/full/455835b.html>



General / reference monolingual corpora (of English) **134**

Last week, tens of thousands of researchers took to the streets to register their opposition to a proposed bill designed to control civil-service spending.



took to the streets

135

- <http://corpus.leeds.ac.uk/internet.html>
- English
- Let's try to understand:
 - Meaning
 - Extended (sentential) co-text, preferential co-selections
 - Context(s) of use
 - Semantic preference
 - Semantic prosody



Using general / reference monolingual corpora (from/on the Web): Leeds Internet corpora

136

Querying Internet corpora

took to the streets
☒ English ([tags](#)) ☐ Chinese ([tags](#)) ☐ Arabic ([tags](#)) ☐ French ([tags](#)) ☐ German ([tags](#)) ☐ Italian ([tags](#)) ☐ Japanese ([tags](#))
☐ Polish ([tags](#)) ☐ Portuguese ([tags](#)) ☐ Russian ([tags&grams](#)) ☐ Spanish ([tags](#))
 ☐ CQP syntax only ([Examples](#)) [Getting help on the query interface](#)



Set parameters of your query

☒ Concordance

Context: (c for characters, w for words)

Sort by: ☐ Document ☐ Frequency ☐ lemma ☒ word
Then by: ☐ left ☒ right

Output: lines

☐ Collocations

Collocation scores: ☐ Mutual Information ☐ T-score ☒ Loglikelihood score

Context: words on the left words on the right

POS tag of the collocate: POS tags

<http://corpus.leeds.ac.uk/internet.html>



Let's explore internal variation

137

- Any other forms of the verb *take*? (colligational constraints)
- Plural/singular of the noun *street*? (colligational constraints)
- Other verbs? (collocational flexibility)
- Other nouns? (collocational flexibility)
- Select “CQP syntax only” *
- [Look at the examples for help](http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPExamples.html)
 - <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPExamples.html>
- Try out a number of different options



Examples of (possible) useful queries

138

- Select the “Concordance” option with the default parameter settings
- Tick (enable) the “CQP syntax only” option (automatic POS-tagging!)
- [lemma="take"] "to" "the" [lemma="street"]
 - *Lemmatised search: finds all possible forms of verb and noun*
- [pos="V.*"] "to" "the" [lemma="street"]
 - *Lemmatised and POS-specific search: as above but finds all verbs*
- [lemma="take"] "to" "the" [pos="N.*"]
 - *Lemmatised and POS-specific search: as above but finds all nouns*
- Click on the “>>” link to the left of the concordance line for context
- To discover which POS (past-of-speech) tags are used in the different corpora for the various languages, click on the associated link “tags”



Now the translation into Italian of “took to the streets” **139**

- Verb?

- andare
- scendere
- ...?

- Noun?

- strada/strade
- piazza/piazze
- ...?

- Preposition?

- in
- nella/nelle
- per la/per le?
- ...?

Which queries do we need? How many are necessary?



Now the translation into Italian of “take to the street” **140**

- [pos="V.*"] [] [lemma="strada"] NB: [] means ‘any word in that position’
- [pos="V.*"] [] [lemma="piazza"]
 - very general (slower)
- [pos="V.*"] [word="(in|nell|nelle)"] [lemma="strada"]
- [pos="V.*"] [word="(in|nell|nelle)"] [lemma="piazza"]
 - more specific/restrictive NB: | is called ‘pipe’, lists alternatives
- [lemma="scendere"] [word="(in|nell|nelle)"] [lemma="strada"]
- [lemma="andare"] [word="(in|nell|nelle)"] [lemma="piazza"]
 - very specific/restrictive



An alternative (for a more complex query)

141

- *Perform a single (more compact but more time-consuming) query:*

[lemma="(scenderelandarelriversarsil...)"] [word="(inlnellalnellel...)"]

[lemma="(stradalpiazzaalvial...)"]

The random selection of returned concordances can give us an indication of the relative frequency of the possible (and attested) combinations, so we can look at their semantic preference and prosody



Monolingual general / reference corpora available online (at least partially, i.e. as demos)

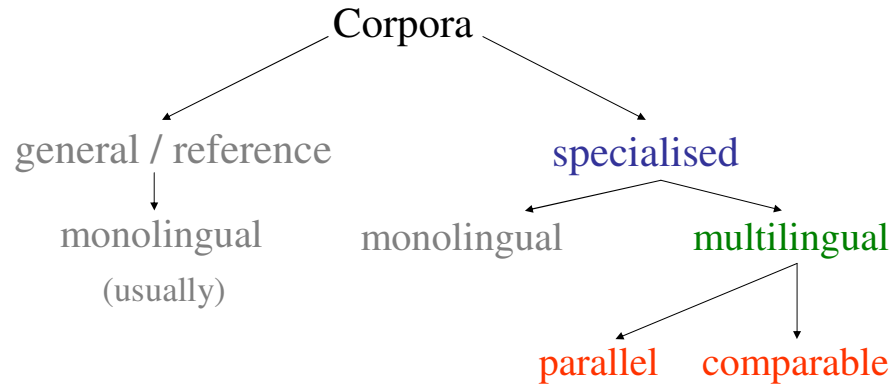
142

- British National Corpus (BNC, British English)
 - <http://www.natcorp.ox.ac.uk> <http://sara.natcorp.ox.ac.uk/lookup.html>
- COCA (American English)
 - <http://www.americancorpus.org>
- La Repubblica corpus (newspaper Italian, approx. 380 M tokens)
 - <http://sslmit.unibo.it/repubblica>
- Leeds Internet corpora
 - English, Chinese, Arabic, French, German, Italian, Japanese, Polish, Portuguese, Russian, Spanish: <http://corpus.leeds.ac.uk/internet.html>
- Mannheim corpora (German)
 - <http://corpora.ids-mannheim.de/ccdb>
- Corpus del Español (Spanish)
 - <http://www.corpusdelespanol.org>
- CREA (Spanish)
 - <http://corpus.rae.es/creanet.html>
- etc. etc., explore the Web to see what other corpora are available!



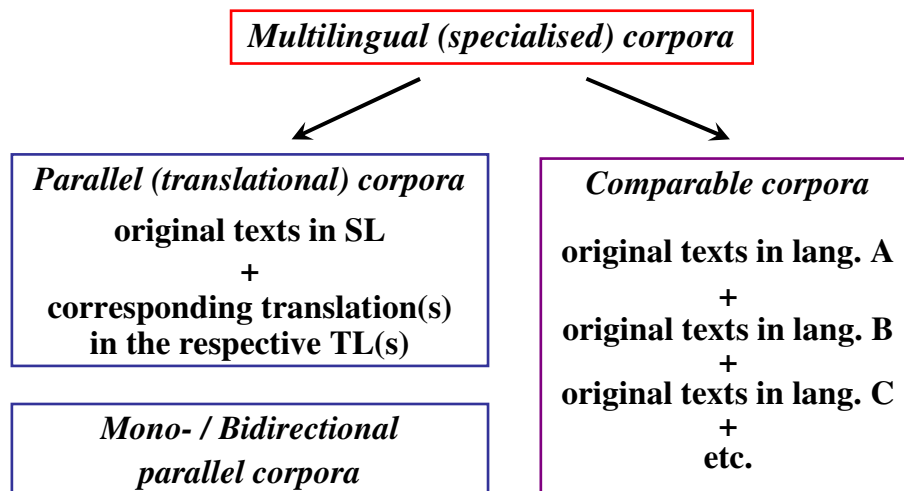
A translation-relevant corpus typology

143



Corpus types for translation (studies)

144



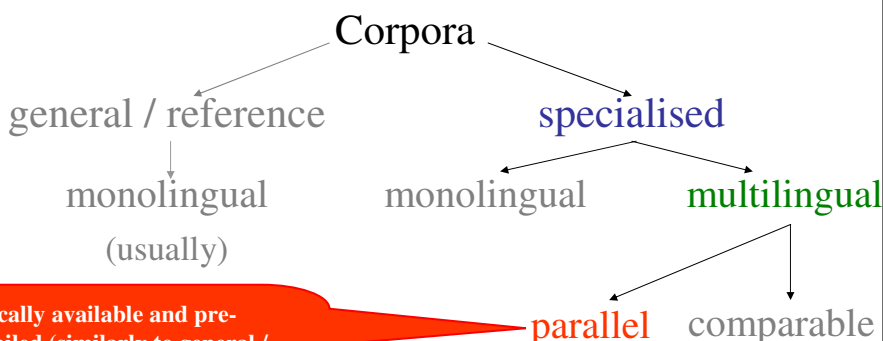


Parallel (translational) corpora

- contain translationally “equivalent” texts: STs and their corresponding TTs
- need to be aligned, usually at the sentence level, i.e. SL sentence X matched to TL sentence X’
- context is provided to account for “equivalence” and “translation shifts” between ST and TT
- translation direction needs to be clear, i.e. which are SL and TL components of the corpus

Comparable corpora

- texts originally produced (**not** translated) in the respective languages
- consist of independent texts which are “similar” according to some pre-determined criteria
- parameters defining this similarity vary widely
- various language components share a set of common features, e.g. text type, genre, publication span, domain, topic



Typically available and pre-compiled (similarly to general / reference monolingual corpora), but it is also possible to build them

Example:

Concordance for the nodeword “issue” (as a noun only) and its Italian counterpart(s) in a corpus of bilingual English/Italian EU parliamentary proceedings (aligned at sentence level), see p. 243



Editorial

Nature **455**, 835-836 (16 October 2008) | doi:10.1038/455835b; Published online 15 October 2008

Cut-throat savings

In an attempt to boost its struggling economy, Italy's government is focusing on easy, but unwise, targets. ▲ Top

It is a dark and angry time for scientists in Italy, faced as they are with a government acting out its own peculiar cost-cutting philosophy. Last week, tens of thousands of researchers took to the streets to register their opposition to a proposed bill designed to control civil-service spending (see page 840). If passed, as expected, the bill would dispose of nearly 2,000 temporary research staff, who are the backbone of the country's grossly understaffed research institutions — and about half of whom had already been selected for permanent jobs.

<http://www.nature.com/nature/journal/v455/n7215/full/455835b.html>



Last week, tens of thousands of researchers took to the streets to register their opposition to a proposed bill designed to control civil-service spending.



*Using bilingual parallel corpora
on the Web: the OPUS corpus*

149

- Going parallel: the OPUS corpus...
 - opus.lingfil.uu.se
- Select:
 - Under “Search & Browse”: *OPUS multilingual search interface*
 - Under “Corpus”: *Europarl3*
 - Under “Languages”: *En*
 - Under “Alignments”: *check it*
- Focus on two separate parts (as basic / initial queries):
 - *control spending*
 - *civil service*



*Using bilingual parallel corpora
on the Web: the OPUS corpus*

150

- These are EU parliamentary debates (proceedings)
- [lem="control"] [] "spending"
 - What equivalents do we find in Italian texts? (set “*show max --- hits*” to 100)
 - For example
 - *contenimento della spesa (pubblica)*
 - *controllare la spesa (pubblica)*
 - *controllo (valido ed efficace) sulla spesa*
 - *controllo della spesa (pubblica)*
 - *rigore della spesa*



*Using bilingual parallel corpora
on the Web: the OPUS corpus*

151

- Now we just focus on the Italian version (without aligned English)
- Which (other) verbs accompany *spesa (pubblica)* and mean “control”?
 - [pos="V.*"] []? "spesa" "pubblica" NB: []? means the slot may be empty
 - (*controllare*), *tagliare*, *ridurre*, *limitare*, *contrarre*, *riorganizzare*
- But... why only [pos="V.*"] ? There is no need for this!
Other possible queries (involving nouns or prepositions)
 - [pos="NOM"] []? "spesa" "pubblica"
 - [pos="PRE"] "spesa" "pubblica"
 - *riduzione*, *controllo*, *contenimento*, *tagli*, *restrizioni*, *limitazione*



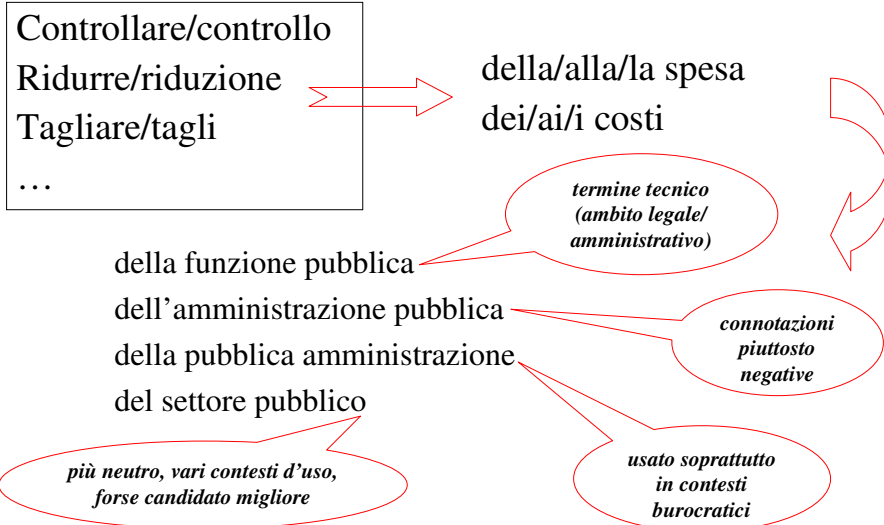
*Using bilingual parallel corpora
on the Web: the OPUS corpus*

152

- ST: *civil-service spending* [vs. *public spending*]
- Let's look for *civil service* in the aligned Eng/Ita texts of the European Parliament with this query: "civil" "service"
- What Italian equivalents do we find?
 - *funzione pubblica*
 - *amministrazione pubblica*
 - *pubblica amministrazione*
 - *settore pubblico*
 - *pubblico impiego*



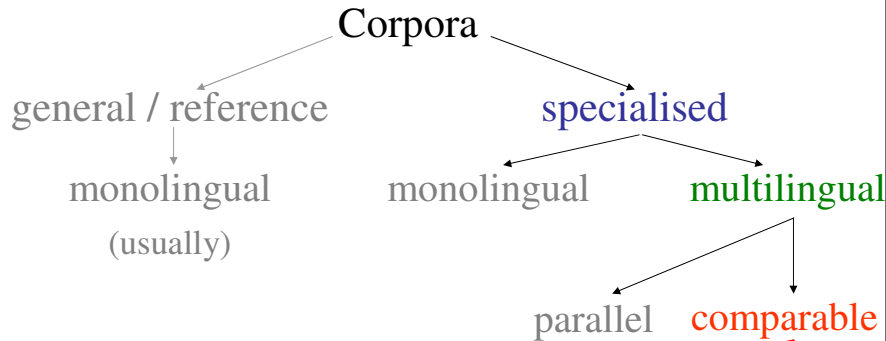
Translation into Italian of “control civil-service spending”¹⁵³
(double-checking context of use in the Italian EU texts)



**Available tools and resources
for parallel (aligned) corpora**

154

- Parallel concordancers
 - Paraconc (demo available for free, but limited functionality)
 - <http://www.athel.com/para.html>
 - Multiconc (old-ish...)
 - http://artsweb.bham.ac.uk/pKing/multiconc/l_text.htm
- Corpora
 - OPUS Corpus: opus.lingfil.uu.se
 - Proceedings of the EU Parliament (EuroParl)
 - Software manuals
 - Subtitling project
 - etc.
 - The texts in the OPUS collection can be consulted/searched online, or downloaded in aligned format



Normally DIY (disposable) corpora, compiled for specific translation assignments, especially for texts belonging to specialised domains



P. R. O. Bally (1959) "Monadenium arborescens". *Candollea* 17:25-26.

Coming from Tanzania, this is a robust growing species and is a semi woody succulent, forming a lightly branched shrub/tree up to 4.25 metres high. The stems can grow to 10 cm. thick, are five angled and may be slightly spirally twisted.

They are erect and may be solitary or in twos. If branched, the branches are quite slender, grow erect, and are some 30 – 60 cm. apart. They are smooth, and covered in a green bloom. Leaf scars, which are 10 mm. in diameter, are borne 4 – 7 cm. apart and below each leaf scar is a small tubercle which on older plants has a small reddish/brown spine, but a more robust one up to 2 cm. long on is produced on younger plants. The leaves are crowded terminally around the ends of the stems and are produced from the angles of the stems. They are obovate, pointed and heart shaped, 7 – 19 cm long and 5.6 – 11 cm. wide. Flowering takes place from an eye situated directly above the leaf scar and several cymes may be produced near the apex of the branches with peduncles 6 – 7 cm. long and 5 – 6 mm. thick. The colour of the inflorescence is red. This species is not in general cultivation due to its rapid growth and size.



The process for manual corpus construction

157

- We want to build a bilingual specialised comparable corpus for a translation task (English => Italian)
- Two stages:
 - a) Source language corpus component (English)
 - b) Target language corpus component (Italian)



Build your own comparable corpus (English / Italian) 158

- Parameters to consider when manually selecting corpus texts
 - topic (relevance)
 - source (authoritativeness)
 - audience (who was it written for)
 - function
 - information-richness
 - language (reliable, accurate, native)
 - date (recent, up-to-date)
 - technical/practical issues, e.g.
 - formatting
 - file type
 - conversion



Practical considerations: file types

159

- Corpus files must be downloaded/saved in this format
 - Simple/pure text (.txt)
 - save as “text only” (different from MS-Word and .doc !!!)
- Common formats of online texts
 - html
 - Microsoft Word
 - pdf
 - image (not good) vs searchable (OK)
 - *these must be converted into (saved as) TXT format*



Searching for similar SL (English) texts for the corpus

160

- We look for:
 - web pages in English, as similar to our ST as possible
 - e.g. searching for ‘monadenium’ on google.co.uk
- We find, e.g.:
 - en.wikipedia.org/wiki/Monadenium_arborescens
 - www.sdcss.com/monadenium.html
 - davesgarden.com/guides/pf/go/65135/
 - www.gardening.eu/plants/Succulent-Plants/Monadenium-guentheri/3708/
 - *You can add to the search string: monadenium filetype:pdf*

In general pages in pdf format tend to be more informative and authoritative

161



Wikipedia is sustained by people like you. Please [donate](#) today.

[Log in](#) / [create account](#)



WIKIPEDIA
The Free Encyclopedia

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox

- Upload file

Monadenium arborescens

From Wikipedia, the free encyclopedia

Monadenium arborescens is a species of [plant](#) in the [Euphorbiaceae](#) family. It is endemic to Tanzania. It is threatened by [habitat loss](#).

Source [edit]

- Carter-Holmes, S. 1998. *Monadenium arborescens*. 2006 IUCN Red List of Threatened Species. [Downloaded on 22 August 2007](#).

 This *euphorb*-related article is a *stub*. You can help Wikipedia by expanding it.

Monadenium arborescens

Conservation status

Extinct Threatened Lower Risk

EX EW CR EN **VU** cd nt lc

Vulnerable (IUCN 2.3)

Scientific classification

Kingdom: [Plantae](#)

Division: [Magnoliophyta](#)

Class: [Magnoliopsida](#)

Order: [Malpighiales](#)

Family: [Euphorbiaceae](#)

Genus: *Monadenium*

Species: ***M. arborescens***

Binomial name

Monadenium arborescens

Bally



Uninformative, different genre

162



[SDCSS Homepage](#)

[Back to Articles](#)

The Genus Monadenium

by Tom De Merritt (Espinosa Y Flores Oct. 2000)

Monadenium belongs in the Euphorbiaceae family, which contains about 320 genera and over 8000 species of geographically widespread plants, which range from annual herbs to large trees. This genus Monadenium has 50 known species, 46 of which are confined to tropical East Africa. Its eastern distribution line is the Indian Ocean. This genus does not exist in Madagascar. Throughout its area of distribution endemic species, varieties and forms of monadenium are numerous, with their habitat often being restricted to a singular hilltop. This high rate of endemism suggests a recent origin and/or a strict response to its environment, which sets limits on the expansion of distribution.

All Euphorbiaceae have a milky sap that may be harmful to the touch. The inflorescence has a complicated structure, based on the cyathium, which consists of an involucre containing one reduced female flower and several male flowers. There are protective bracts and nectarines. The fruit is a capsule, which explodes on reaching maturity.

The word "Monadenium" is derived from Greek, meaning "Single Gland". F. Pax originated the name in 1895 based on the initial species of the genus to be described, *Monadenium coccineum*. This distinguishing characteristic for which the name was applied emanates from the flower structure. Flowers consist of a cup-like bracteate head, truncate at the top, open at one side or below the middle with a "single gland" of varying



Very informative, authoritative (source: San Diego Cactus Society), similar genre (journal article)


163

Dave's Garden

"For Gardeners... By Gardeners"

[Home](#) | [My Tools](#) | [Communities](#) | [Guides & Information](#) | [Products & Sources](#) | [About Us](#)

[Search and Browse Tools](#) | [Add a plant](#) | [Zone lookup](#) | [F.A.Q.](#) | [Advertise](#) | [Register](#) | [Login](#)

Oct 17, 7:14 AM

PlantFiles: Monadenium

Monadenium ritchei

Guides and Information

- Articles
- Videos
- PlantFiles
- Botany
- BugFiles
- Insectipedia
- Garden Terms
- Landscaping

Family: [Euphorbiaceae](#) (yoo-for-bee-AY-see-ee) (Info)

Genus: [Monadenium](#) (mon-uh-DEE-nee-um) (Info)

Species: [ritchey](#)

2 members have or want this plant for trade.

Category:
Cactus and Succulents

Height:
6-12 in. (15-30 cm)
12-18 in. (30-45 cm)

Spacing:
12-15 in. (30-38 cm)

Hardiness:
USDA Zone 11: above 4.5 °C (40 °F)



Uninformative, little connected text, different function (promotional) and genre


164

Gardening!eu

[Home](#) | [My Tools](#) | [Communities](#) | [Guides & Information](#) | [Products & Sources](#) | [About Us](#)

[Search and Browse Tools](#) | [Add a plant](#) | [Zone lookup](#) | [F.A.Q.](#) | [Advertise](#) | [Register](#) | [Login](#)

Oct 17, 7:14 AM

Monadenium guentheri

garden

<%login%>

Find plants

index

%=parole("Sezioni")%:

Apartment

Orchids

Shrubs

Trees

Annual

Climbers

Aquatic

Perennial

Bulbs

Garden

Euphorbiaceae

Fertilization Monadenium guentheri

In this period of the year, let's remember to enrich the soil of our succulent plants, using, every 15-20 days, a specific fertilizer: fertilizers for succulent plants must be poor in nitrogen to avoid them from producing too much vegetation which would be easily attacked by fungal diseases.

Generalities Monadenium guentheri

These plants are succulent. Monadenium guentheri is an evergreen; during winter it assumes a greenpink colouring; the adult specimens are small in size and reach 30 cm high. When growing these plants tend to go up, developing very little in width, and they tend to have a columnar development.

Watering Monadenium guentheri

Let's remember that the succulent plants can withhold water inside their materials, therefore they need to be watered only when the weather is hot. Let's avoid watering the The Monadenium guentheri excessively, always leaving the soil dry for a few days between one watering and the other, therefore let's indicate the substitution



Low quality, unreliable (language)



Build the SL (English) corpus component manually 165

- When you find online texts that you want to include in the corpus
- From the browser
 - html file
 - File=> save as=> filename.txt
 - If possible, try not to modify the file name, just the file extension
 - Save as (format)=> text file
 - pdf file (searchable)
 - Edit=> select all=> copy => paste in a text file => save
 - MS Word file
 - Save as filename.txt
 - File type> “plain text *.txt” => ok (ignore any error messages)
- Plan separate folders for each corpus (sub-)component
 - e.g. SL/TL, but also more/less authoritative (due to language or source)



Main sources 166

- Austermühl, F. (2001) *Electronic Tools for Translators*. Manchester, St. Jerome Publishing: Chapter 8 “Corpora as Translation Tools” (pages 124-133)
- Bowker, L. (2002) *Computer-Aided Translation Technology. A Practical Introduction*. Ottawa, University of Ottawa Press: Chapter 2 “Capturing Data in Electronic Form” (pages 22-42) & Chapter 3 “Corpora and Corpus-Analysis Tools” (pages 43-76)
- Laviosa, S. (2003) “Corpora and the translator”: Chapter 7 in H. Somers (ed.) *Computers and Translation. A Translator’s Guide*. Amsterdam and Philadelphia, John Benjamins (pages 105-117)
- Bowker, L. & J. Pearson (2002) *Working with Specialized Language: a Practical Guide to Using Corpora*. London and New York, Routledge: this book is recommended as a particularly accessible and comprehensive introduction to the use of corpora



Further (optional) reading

167

These additional books provide a more technical and in-depth discussion of how corpora can be built (e.g. from the Web) and used, for students who are interested:

- Baroni, M. & S. Bernardini (eds) (2006) *Wacky! Working Papers on the Web as Corpus*. Bologna: GEDIT. Freely downloadable from <http://wackybook.sslmit.unibo.it/>
- Gatto, M. (2009) *From Body to Web: An Introduction to the Web as Corpus*. Bari/Roma: Università di Bari/Editori Laterza. Downloadable after free registration from <http://www.universitypressonline.it/pubblicazioni>
- McEnery, T., R. Xiao & Y. Tono (2006) *Corpus-based Language Studies: an Advanced Resource Book*. London and New York: Routledge
- Zanettin, F. (2012) *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome



Further (optional) reading

168

- From the MeLLANGE online materials and e-learning resources (see slides 94-95) select the topic “Corpora for Translation”
- This e-learning course on (the use of) corpora for translation is broken down into 5 self-contained units (or lessons) which can be completed in a step-by-step fashion and contain explanations, tests, quizzes, exercises, etc.
- Use these e-learning materials on corpora for translation to consolidate what was covered in class, and to familiarise yourself with the English terminology which is required to discuss corpora (for the exam!)



***Build the TL corpus component (Italian)
(semi-)automatically with BootCaT***

169

- <http://bootcat.sslmit.unibo.it>
 - BootCaT is a free tool to build corpora (semi-) automatically, rather than manually, from the Web
 - Install BootCaT on your computer (including Strawberry Perl):
 - <http://bootcat.sslmit.unibo.it/?section=download>
 - Study the tutorial on how to use BootCaT's front-end:
 - http://docs.sslmit.unibo.it/doku.php?id=bootcat:tutorials:basic_1
 - To use BootCaT, you need to get a (free!*) Bing AppId:
 - http://docs.sslmit.unibo.it/doku.php?id=bootcat:help:search_engine_key
- * Limited to 5,000 queries/tuples per month



***Extract from a botany corpus (in English)
built with BootCaT***

170

is a genus of approximately 170 species [1] commonly called dayflowers due to the short lives of their flowers. They are less often known as widow's tears . It is by far the largest genus of its family, Commelinaceae . [2] The Swedish taxonomist Carl Linnaeus of the 18th century named the genus after the two Dutch botanists Jan Commelin and his nephew and Kaspar, each representing one of the showy petals of *Commelina communis* . [1] The dayflowers are herbs that may be either perennial or annual. They are characterised by their zygomorphic flowers and by involucre bracts called spathes that surround the flower stalks. [3] These spathes are often filled with a mucilaginous liquid. Each spathe houses either 1 or 2 scorpioid cymes , with the upper cyme being either vestigial or bearing 1 to several typically male flowers, and the lower cyme bearing several flowers. All members of the genus have alternate leaves. [1] The Asiatic dayflower (*Commelina communis*) is probably the best known species in the West. It is a common weed in parts of Europe and throughout eastern North America [1] . Several species such as *Commelina benghalensis* , are eaten as a leaf vegetable in Southeast Asia and Africa . [4] [edit] Description Plants in the genus are perennial or annual herbs with roots that are usually fibrous or rarely tuberous or rhizomatous . The leaves are distichous (i.e. 2-ranked) or spirally arranged with blades that either lack or have a petiole. The ptyxis, or the way the leaf is folded in the bud, is either involute (i.e. having rolled margins) or supervolute. [5] [1] The inflorescences are terminal, meaning the stem terminates with an inflorescence, and often leaf opposed, meaning it emerges at the node with a leaf of a new axillary stem. The inflorescence is composed of 1 or 2 cincinni, also called scorpioid cymes , which are monochasiums (i.e. cymes with a single branched main false axis) in which the lateral branches arise alternately on opposite sides of the false axis. The distal cincinnus may either be vestigial or contain one to several flowers that are typically male. The proximal cyme is always present and is several flowered. The cincinni are enclosed in a folded spathe modified leaf, which is often filled with a mucilaginous liquid. The spathe may either have completely distal margins or they may be fused to varying degrees at the basal end. [5] [1] The flowers are borne on pedicels as



*Now that we have built the corpus
what concordancing tools are available?*

171

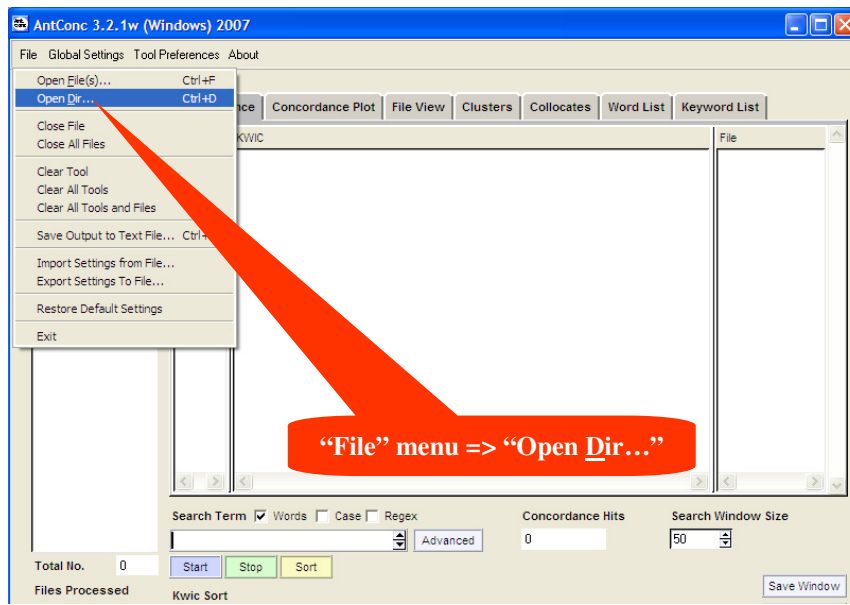
- For example
 - *AntConc* - I used this software for the slides (with a corpus created semi-automatically with BootCaT), and you can download it (for free and legally!!) from this URL:
 - <http://www.antlab.sci.waseda.ac.jp/software.html>
 - *TextStat*:
 - <http://www.niederlandistik.fu-berlin.de/textstat/software-en.html>

And what can we do with them?



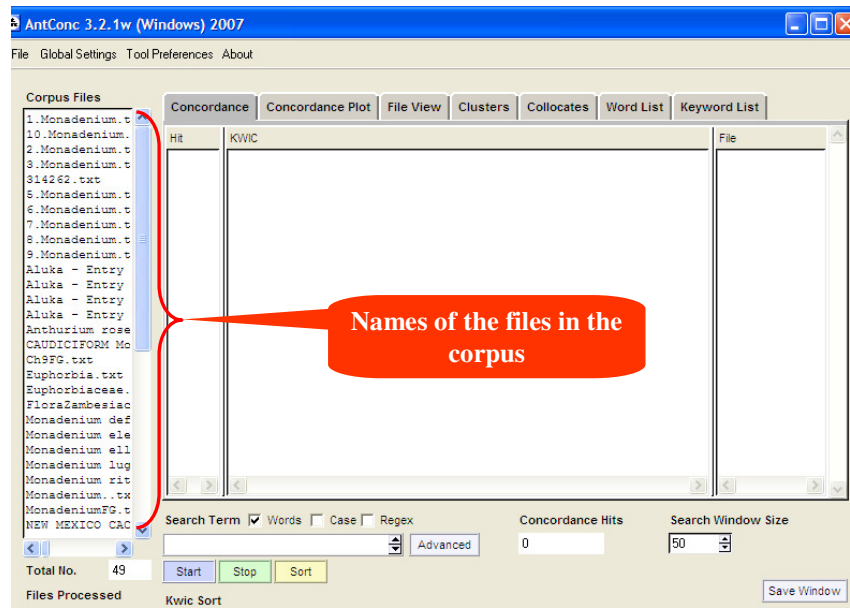
AntConc

172





173



Find or investigate collocates from a comparable Eng/Ita corpus on botany (with AntConc)

174

- search the English corpus for collocates of:
 - *roots*
 - in a span (i.e. distance/space) of 2L/1L
 - with min. fq of collocate=2
- then search a comparable Italian corpus for collocates of:
 - *radici*
 - in a span (i.e. distance/space) of 1R/2R
 - with min. fq of collocate=2

175



AntConc 3.2.1w (Windows) 2007

File Global Settings Tool Preferences About

Corpus Files

onadenius_en_raw

Concordance Concordance Plot File View Clusters Collocates Word List Keyword List

Total No. of Collocate Types: 42 Total No. of Collocate Tokens: 182

Rank	Freq	Freq(L)	Freq(R)	Collocate
1	16	16	0	the
2	13	13	0	of
3	13	13	0	herbs
4	10	10	0	from
5	9	9	0	fleshy
6	7	7	0	À
7	7	7	0	with
8	7	7	0	fibrous
9	7	7	0	and
10	6	6	0	The
11	6	6	0	or
12	6	6	0	Climbing
13	5	5	0	tuberous
14	5	5	0	fusiform
15	4	4	0	rot
16	4	4	0	on
17	3	3	0	thickened

Search Term ☒ Words ☐ Case ☐ Regex

From... 2L To... 1L

Min. Collocate Frequency 2

Sort by Sort by Freq

Files Processed 1

Reset

Save Window

Exit

176



AntConc 3.2.1w (Windows) 2007

File Global Settings Tool Preferences About

Corpus Files

ca_cactus_corpus

Concordance Concordance Plot File View Clusters Collocates Word List Keyword List

Total No. of Collocate Types: 107 Total No. of Collocate Tokens: 466

Rank	Freq	Freq(L)	Freq(R)	Collocate
1	36	0	36	e
2	26	0	26	che
3	18	0	18	di
4	17	0	17	si
5	13	0	13	sono
6	13	0	13	o
7	10	0	10	avventizie
8	9	0	9	non
9	8	0	8	À
10	8	0	8	a
11	7	0	7	lunghe
12	7	0	7	il
13	7	0	7	fibrose
14	7	0	7	carnose
15	6	0	6	un
16	6	0	6	tuberose
17	6	0	6	possono

Search Term ☒ Words ☐ Case ☐ Regex

From... 1R To... 2R

Min. Collocate Frequency 2

Sort by Sort by Freq

Files Processed 1

Reset

Save Window

Exit



177

Concordance				
Concordance Plot		File View		Clusters
Total No. of Collocate Types: 42		Total No. of Collocate Tokens: 182		
Rank	Freq	Freq(L)	Freq(R)	Collocate
1	16	16	0	the
2	13	13	0	of
3	13	13	0	herbs
4	10	10	0	from
5	9	9	0	fleshy
6	7	7	0	à
7	7	7	0	with
8	7	7	0	fibrous
9	7	7	0	and
10	6	6	0	The
11	6	6	0	or
12	6	6	0	climbing
13	5	5	0	tuberous
14	5	5	0	fusiform
15	4	4	0	rot
16	4	4	0	on
17	3	3	0	thickened
18	2	2	0	stems

Concordance				
Concordance Plot		File View		Clusters
Total No. of Collocate Types: 107		Total No. of Collocate Tokens: 486		
Rank	Freq	Freq(L)	Freq(R)	Collocate
1	36	0	36	e
2	26	0	26	che
3	18	0	18	di
4	17	0	17	si
5	13	0	13	sono
6	13	0	13	o
7	10	0	10	avventizie
8	9	0	9	non
9	8	0	8	à
10	8	0	8	a
11	7	0	7	lunghe
12	7	0	7	il
13	7	0	7	fibrose
14	7	0	7	carnose
15	6	0	6	un
16	6	0	6	tuberose
17	6	0	6	possono
18	6	0	6	...

Corpus query tool: AntConc



Study concordances in a comparable Eng/Ita corpus on botany (with AntConc)

178

- first search the English corpus for leaves are
- sort concordance results according to the 2nd word to the right (2R)
- then search the Italian corpus for foglie sono
- sort concordance results in the same way

179



AntConc 3.2.1w (Windows) 2007

File Global Settings Tool Preferences About

Corpus Files

monadenium_en_ra

Concordance

Hit	KWIC	File
1	great and the fuzzy white leaves are a treasure. It is quite o	monadenium_en_ra
2	ra. Usually only the lower leaves are affected. Pathogen and dis	monadenium_en_ra
3	e of the edible roots. The leaves are also edible and are usefu	monadenium_en_ra
4	ne species is a tree). The leaves are alternate and variable and	monadenium_en_ra
5	an ash-coloured bark. The leaves are alternate, three centimet	monadenium_en_ra
6	ved. The large, soft, oval leaves are alternate or opposite whe	monadenium_en_ra
7	umerous upright stems, the leaves are alternate, broadly ovate	monadenium_en_ra
8	ing or trailing stems. The leaves are alternate, ovate and entil	monadenium_en_ra
9	e and white edged crimson. Leaves are alternate or in simple sp	monadenium_en_ra
10	rrel. A Celastrus scandens leaves are alternate, dark green, ov	monadenium_en_ra
11	all In Arabia and Java the leaves are applied to skin complaints	monadenium_en_ra
12	an U. alpina allspathulate leaves are at most only 7 cm long and	monadenium_en_ra
13	touches the soil surface. Leaves are borne on terminal growth;	monadenium_en_ra
14	ia Building larger shrubs. Leaves are dark red to purple when y	monadenium_en_ra
15	long, ascending and rigid. Leaves are deciduous, up to 40 mm lo	monadenium_en_ra
16	with alternate leaf scars. Leaves are deciduous, crowded on the	monadenium_en_ra
17	k (the medicinal part). A Leaves are deciduous, alternate, spi	monadenium_en_ra
18	t 33 meters in length. The leaves are deep green, divided into	monadenium_en_ra
19	erous or rhizomatous. The leaves are distichous (i.e. 2-ranked	monadenium_en_ra
20	with age. The olive-green leaves are distinctly long and narrow	monadenium_en_ra
21	bulbuls and starlings. The leaves are eaten by animals includin	monadenium_en_ra
22	s ornamental. In Java the leaves are eaten as vegetable (Burbil	monadenium_en_ra

Search Term ☒ Words ☐ Case ☐ Regex

leaves are

Concordance Hits 51

Search Window Size 50

Total No. 1

Files Processed

Reset

Kwic Sort

☒ Level 1 ☐ Level 2 ☐ Level 3

Start Stop Sort

Save Window

Exit

180



AntConc 3.2.1w (Windows) 2007

File Global Settings Tool Preferences About

Corpus Files

ita_cactus_corpu

Concordance

Hit	KWIC	File
1	e a radicare ai nodi. Le foglie sono alterne, spatolate o spat	ita_cactus_corpu
2	resce in luoghi piani. Le foglie sono alterne, rotondato-lobate,	ita_cactus_corpu
3	in senso orizzontale. Le foglie sono alterne, astate, con cont	ita_cactus_corpu
4	striato e pubescente. Le foglie sono alterne, bi-tripennate, d	ita_cactus_corpu
5	igliano sono levogiri. Le foglie sono alterne, le inferior... du	ita_cactus_corpu
6	crecita delle lamine. Le foglie sono alterne, con lamine lineari	ita_cactus_corpu
7	to ramificato e denso; le foglie sono arrotondate, curiose, di c	ita_cactus_corpu
8	e specie mediterranee, le foglie sono assenti o precocemente cad	ita_cactus_corpu
9	bianchi. Anche le strette foglie sono bianche e lanuginose, spe	ita_cactus_corpu
10	singoli fusti sottili, le foglie sono c... dettagli di: Fragaria	ita_cactus_corpu
11	mensioni di un albero. Le foglie sono caduche, ma molte specie	ita_cactus_corpu
12	di diverse dimensioni. Le foglie sono carnosse e di colore verde,	ita_cactus_corpu
13	sciolati sugli spigoli; le foglie sono carnosse e presto caduche,	ita_cactus_corpu
14	lià Maioe arboreo-scens. Le foglie sono carnosse e protette da spin	ita_cactus_corpu
15	semplice o ramificata. Le foglie sono carnosse, più o meno denti	ita_cactus_corpu
16	altra prive di fusto. Le foglie sono carnosse, coniche o piramid	ita_cactus_corpu
17	grigia o rosso-bruna. Le foglie sono composte da 9-30 fogliolli	ita_cactus_corpu
18	otteccia l' rossastra; le foglie sono composte, costituite da li	ita_cactus_corpu
19	veria elegans all le sue foglie sono concave e ricoperte di una	ita_cactus_corpu
20	portamento prostrato, le foglie sono coperte da una fitta pelu	ita_cactus_corpu
21	un bel fogliame denso. Le foglie sono cuoriformi e rugose, di co	ita_cactus_corpu
22	del mimetismo. In cui le foglie sono del tutto assenti e sostit	ita_cactus_corpu

Search Term ☒ Words ☐ Case ☐ Regex

foglie sono

Concordance Hits 117

Search Window Size 50

Total No. 1

Files Processed

Reset

Kwic Sort

☒ Level 1 ☐ Level 2 ☐ Level 3

Start Stop Sort

Save Window

Exit



Study concordances in a comparable Eng/Ita corpus on botany

181

ance	Concordance Plot	File View	Clusters	Collocates	Word List	Keywo	ance	Concordance Plot	File View	Clusters	Collocates	Word List	Keywo
KVMC great and the fuzzy white leaves are a treasure. It is quite o: rm. Usually only the lower leaves are affected. Pathogen and di: e of the edible roots. The leaves are also edible and are usefu: ne species is a tree). The leaves are alternate and variable and an ash-coloured bark . The leaves are alternate, three centimet: ved. The large, soft, oval leaves are alternate or opposite when umerous upright stems, the leaves are alternate, broadly ovate : ling or trailing stems. The leaves are alternate, ovate and entir: e and white edged crimson. Leaves are alternate or in simple sp: rrel.À Celastrus scandens leaves are alternate, dark green, ovi: all In Arabia and Java the leaves are applied to skin complaints an U. alpina allspathulate leaves are at most only 7 cm long and touches the soil surface. Leaves are borne on terminal growth; ia Building larger shrubs. Leaves are dark red to purple when y: long, ascending and rigid. Leaves are deciduous, up to 40 mm lo: with alternate leaf scars. Leaves are deciduous, crowded on the k (the medicinal part). À Leaves are deciduous, alternate, spic: t 33 meters in length. The leaves are deep green, divided into : erous or rhizomatous . The leaves are distichous (i.e. 2-ranked; with age. The olive-green leaves are distinctly long and narrow bulbous and starlings. The leaves are eaten by animals includin: s ornamental. In Java the leaves are eaten as seasoning (Burkil)							KVMC e a radicare ai nodi. Le foglie sono alterne, spatolate o spat: resce in luoghi piani. Le foglie sono alterne, rotondato-lobate. in senso orizzontale. Le foglie sono alterne, astate, con cont: striato e pubescente. Le foglie sono alterne, bi-tripennate, d: igliano sono levogiri. Le foglie sono alterne, le inferior... de crescita delle lamine. Le foglie sono alterne, con lamine lineal: to ramificato e denso; le foglie sono arrotondate, curiose, di (: e specie mediterranee, le foglie sono assenti o precocemente cad: bianchi. Anche le strette foglie sono bianche e lanuginose, spec: singoli fusti sottili, le foglie sono c... dettagli di: Fragaria mensioni di un albero. Le foglie sono caduche, ma molte specie : di diverse dimensioni. Le foglie sono carnose e di colore verde. rcolati sugli spigoli; le foglie sono carnose e presto caduche. lallaloe arborescens. Le foglie sono carnose e protette da spin: semplice o ramificata. Le foglie sono carnose, più o meno dent: altra prive di fusto. Le foglie sono carnose, coniche o piramid: grigia o rosso-bruna. Le foglie sono composte da 9-30 fogliolli: orteccia À rossastra; le foglie sono composte, costituite da l(everia elegans all le sue foglie sono concave e ricoperte di una portamento prostrato, le foglie sono coperte da una fitta pelu: un bel fogliame denso. Le foglie sono cuoriformi e rugose, di c: del munitono in cui le foglie sono del tutto assenti e sostit						



Using a comparable corpus: understand (parts of) the source text

182

- Example: “robust”

P. R. O. Bally (1959) “Monadenium arborescens”. *Candollea* 17:25-26.

Coming from Tanzania, this is a **robust** **growing species** and is a semi woody succulent, forming a lightly branched shrub/tree up to 4.25 metres high. The stems can grow to 10 cm. thick, are five angled and may be slightly spirally twisted.

183

- Write 'robust' in the "search term" box
- Make sure the box next to "Words" is ticked
- Click on 'Start'

The file in which the term was found

Search term in context (4-5 words to the left and right)

Search term

Tick this option to use regular expressions

184

- Ordering concordance lines: "Kwic Sort" => tick "Level 1"
- Scroll down to 1R (1 word to the right)
- Click on "Sort"
- The words immediately to the right of the search term are now red and have been grouped together (*robust growing/growth*)

Sort



*Using a comparable corpus:
now turn to the TL to produce the actual translation*

185

- When we are sure that we understand the ST, we make hypotheses as to possible translation equivalents for “robust growing species”
- We close the SL corpus (in English in our case)
 - menu file=> “close all files”
- We open the TL corpus (in Italian in our case)
 - menu file=> “open dir”
- We repeat the steps followed for English, searching for evidence supporting our hypotheses
- The comparable corpus helps us to replicate the style of domain experts in the TL (terms, phraseology...)



Comparing corpora and the Web

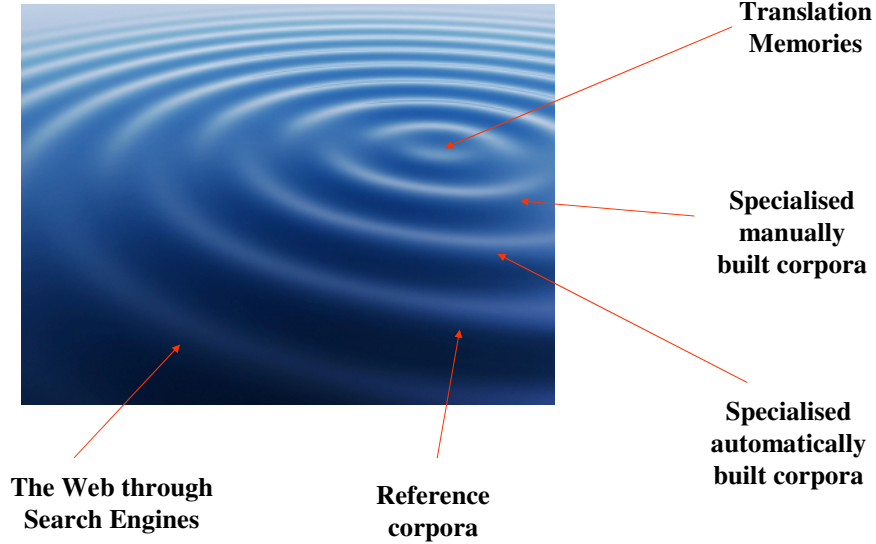
186

- Compare capabilities seen here using corpora to our experience of using Google for linguistic queries
- Corpora:
 - Usually stable
 - searches can be replicated
 - Control over contents
 - we can select the texts to be included
 - Ad-hoc linguistically-aware software to investigate them
 - concordancers (e.g. *AntConc*) can sort / organise concordance lines
 - *regular expressions* possible (ask me if you'd like to learn more about them!)
- Web (as accessed via Google or other search engines):
 - Very unstable
 - results can change at any time for reasons beyond our control
 - No control over contents
 - what/how many texts are indexed by Google's robots?
 - Limited control over search results
 - cannot sort or organise hits meaningfully; they are presented randomly



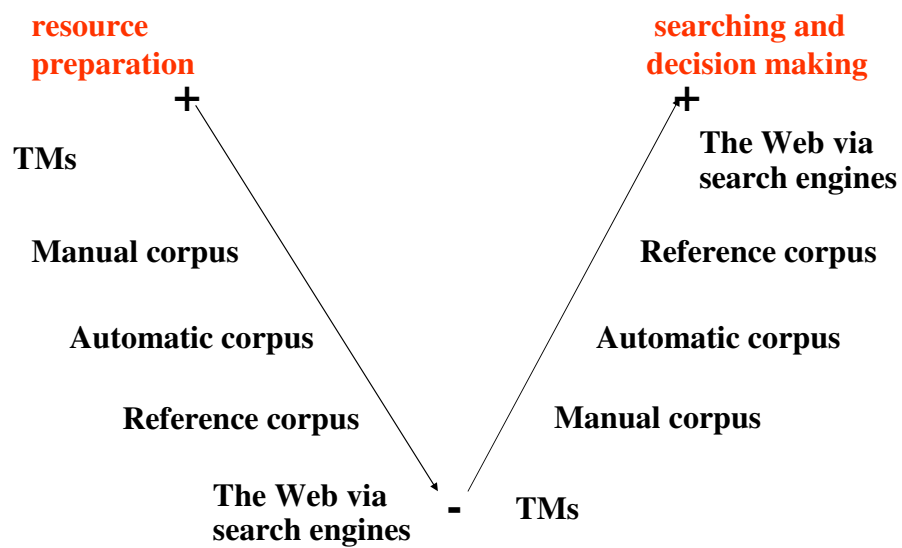
Concluding remarks

187



Time and effort required for...

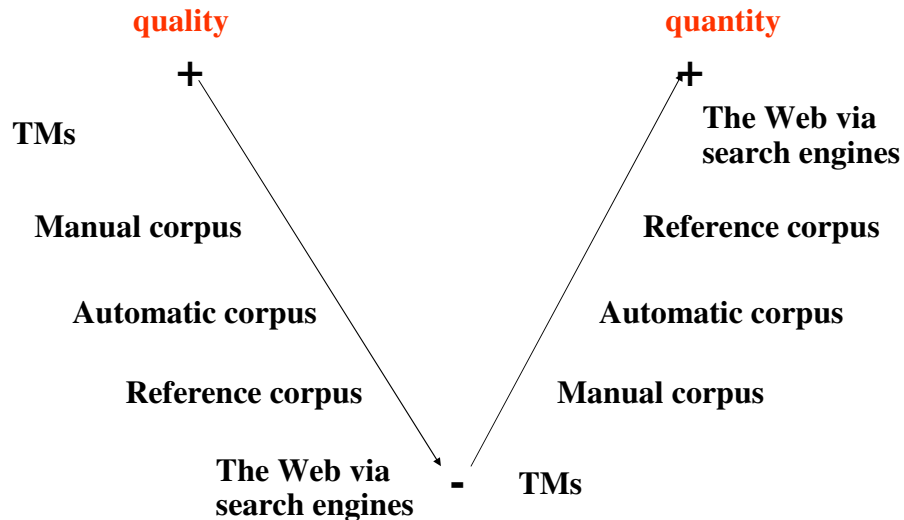
188





Quality vs. quantity

189



NB: how much of the quantity is actually relevant to our translation needs?



Rules for the written exams of modules A and B (Academic Year 2012/2013)

190

- Separate written exams for module A (Prof. Di Giovanni) and B (Prof. Gaspari)
- Each exam counts 50% towards the overall/final mark for the whole course
- You can take the 2 written exams for the 2 modules at any time in any order
- You may take the two exams in the same *appello/sessione* or in different ones
- Frequency of written exams (available *appelli* throughout the academic year):

<u>Exam session</u>	<u>Number of appelli for each module</u>
Summer (2013)	2
Autumn (2013)	1
Winter (2014)	1

- You can refuse a pass mark for an exam (of either module), and then take the same exam again at another *appello*; the refused mark is no longer valid



Format of the exam for module B **191**
“Strumenti e tecnologie per la traduzione specialistica”

- 2-hour written test
- Only monolingual English dictionary is allowed, nothing else
- Answer any 2 of 4 questions; all questions are equally weighted
- Each of the available questions (in English) focuses on a different area/ topic covered in the module
- Answers must be in English
- Expected length is approximately 300 words for each answer
- Marking and assessment will be based on:
 - technical correctness and relevance of the answers (respecting word limit)
 - overall quality of the language used (grammar, vocabulary, syntax, clarity, fluency, structure of the answers, etc.)



General advice for the exam for module B **192**

- Make sure that your answers are relevant to the specific question
- Choose wisely which of the 2 questions (out of 4) you answer: they should be the ones for which you can give the best answers
- If you think that this is helpful, feel free to illustrate the points that you make with suitable examples involving English and Italian, and/or referring to translation instances with these languages in either direction
- Make an effort to come up with your own relevant examples, avoid reusing those discussed in class, on the slides, in your books, etc.
- You are not expected to provide verbatim quotations from sources, but are encouraged to use your own words in answering the questions (e.g. for definitions and explanations of concepts)



Sample exam questions for module B

193

- Explain the concepts of X, Y and Z
- Discuss some of the main strengths and weaknesses associated with using X for specialised translation
- Give an overview of the types of X that can be used by translators working on specialised texts
- Explain the notion of X and discuss its importance for translation
- Explain the differences between X and Y as resources (or tools) that can be used for specialised translation
- Discuss in what circumstances, how and for what purposes X can be used by professional translators



Sample exam questions for module B

194

- Explain the differences and similarities between the notions of X and Y with reference to the development and use of translation tool Z
- What are the main benefits and advantages of using translation tool X from the point of view of professional translators?
- Discuss some of the main problems and limitations involved in relying on X as a resource/tool that can help translators working on specialised texts
- Explain how X and Y typically work and how they can be profitably used by professional translators dealing with specialised texts
- Give an overview of the ways in which X (and Y) can support specialised translation
- Describe a realistic scenario in which using X would be particularly useful for a translator, and explain why this is the case