

Developing linguistic resources from corpus material — a localisation perspective —



Harold Somers
*Centre for Computational Linguistics,
UMIST, Manchester,
England*



SCALLA 2001, NCST Bangalore

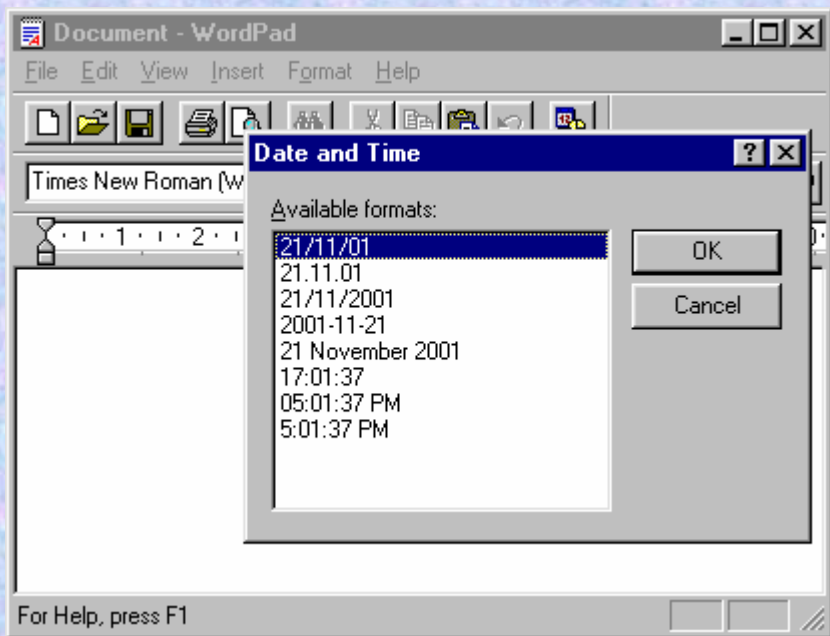
What is localisation?

- “Localisation involves taking a product and making it linguistically, technically, and culturally appropriate to the target locale where it will be used and sold.” (LISA)
- Internationalisation
- Globalisation

Source: Bert Esselink chapter in Somers (ed.) (forthcoming)

What gets “localised”?

- Software, including
 - Messages, dialogues, menus (incl. hot keys)



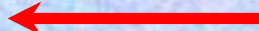
What gets “localised”?

- **Software, including**
 - Messages, dialogues, menus (incl. hot keys)
 - sample files
 - online help
 - online and printed documentation
 - collateral materials
 - multimedia demos
- **Web pages**
- **Other products**

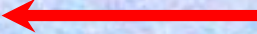
Aspects of localisation

- Language
- Culture
- Regional standards such as
 - character set
 - Currency
 - default page sizes
 - address formats
 - custom calendars
 - date/time formats.

Project process

- Project setup
- Translation 
- Review
- Production
- Quality assurance
- Project closure

Tools and resources needed

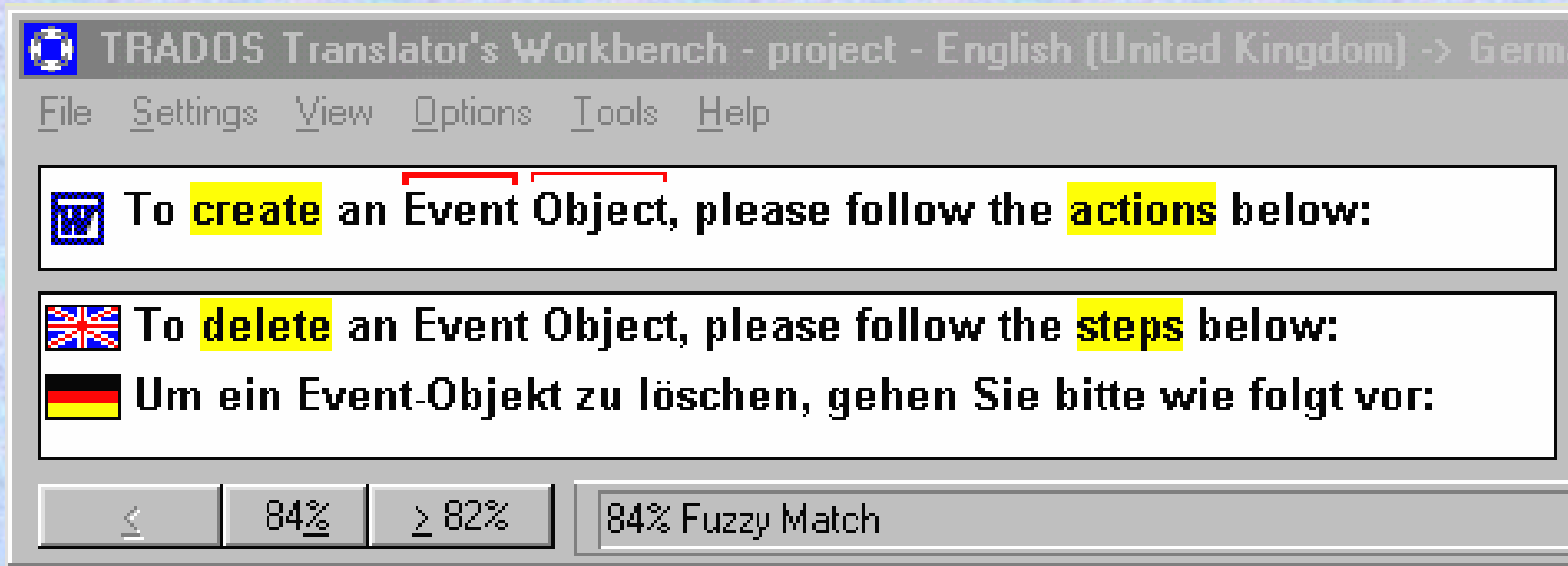
- Computer-aided translation tools:
 - Word processing tools
 - Online dictionaries
 - Terminology
 - Translation memory 

Translation Memory

- Definition
- Brief history
- Main features
- Use of corpus-based techniques
- Translation Memory and EBMT

What is Translation Memory?

- Database of previously translated pairs of equivalent source-/target-language segments (typically, sentences)
- Software to search the database



Brief history

- Kay 1980 often cited as originator, ...
- but Peter Arthern (1978) made clear suggestion, and Alan Melby's group was working on the idea in 1970s
- “Repetitions processing” part of ALPS system in 1981
- current term used (and technology widely available) from early 1990s

Main features

- Database of previous translations
 - Alignment tool
- Efficient mechanism for retrieving candidate matches
 - “fuzzy matching” component
 - linguistically sophisticated matching?
- Terminology look-up and other features
- User must decide usefulness ...
- ... and *how* to use matches

Fuzzy matching

a. When the paper tray is empty, remove it and refill it with paper of the appropriate size.

b. When the tray is empty, remove it and
..... fill it with the appropriate paper.

c. When ~~the bulb remains unlit~~, remove it
~~and replace it with a new~~ bulb.

~~d. You have to remove the paper tray in order
to refill it when it is empty.~~

Pattern match

Structural match

Semantic match

Database of previous translations

- Where do we get these from?
 - Similar texts?
 - Previously localised software (circular)
 - Possible use of comparable corpora
 - Build TM as you go
- What do we do with them?
 - Extract terminology
 - Align them at word and sentence level

also useful as concordance

Document Collection: **Canadian Hansard (1986-1993)**

Expression: **point of order**

- | | | |
|-----|--|---|
| 1. | «Monsieur le Président, j'invoque le Règlement.» | “Mr. Speaker, point of order ”. |
| 2. | À l'époque, on pouvait invoquer le Règlement même pendant la période des questions. | In those days one could interrupt anything with a point of order , even Question Period. |
| 3. | Monsieur le Président, j'invoque le Règlement. | Mr. Speaker, I rise on a point of order . |
| 4. | J'invoque le Règlement, monsieur le Président. | Mr. Speaker, I rise on a point of order . |
| 5. | Monsieur le Président, j'invoque le Règlement. | Mr. Speaker, I rise on a point of order . |
| 6. | Monsieur le Président, j'invoque le Règlement. | Mr. Speaker, I rise on a point of order . |
| 7. | Avant de passer la parole au député de Moncton, je passe la parole au secrétaire parlementaire du leader du gouvernement à la Chambre pour un rappel au Règlement. | Before I recognize the hon. member for Moncton I have a point of order from the hon. parliamentary secretary to the government House leader. |
| 8. | Pour un rappel au Règlement, monsieur le Président. | I rise on a point of order , Mr. Speaker. |
| 9. | Monsieur le Président, j'invoque le Règlement. | Mr. Speaker, a point of order . |
| 10. | J'invoque le Règlement, monsieur le Président, pour déclarer que ce que vient de dire le député est totalement et absolument faux. | On a point of order , Mr. Speaker. I want to say that the point the member just made is absolutely and totally false. |

Source: TransSearch
system

developed by RALI,
Université de Montréal.

Alignment

- Explicit juxtaposition of corresponding portions of parallel text
 - At various levels, typically paragraph and sentence, if appropriate
 - Word-level alignment may be a side-effect or a goal

Techniques

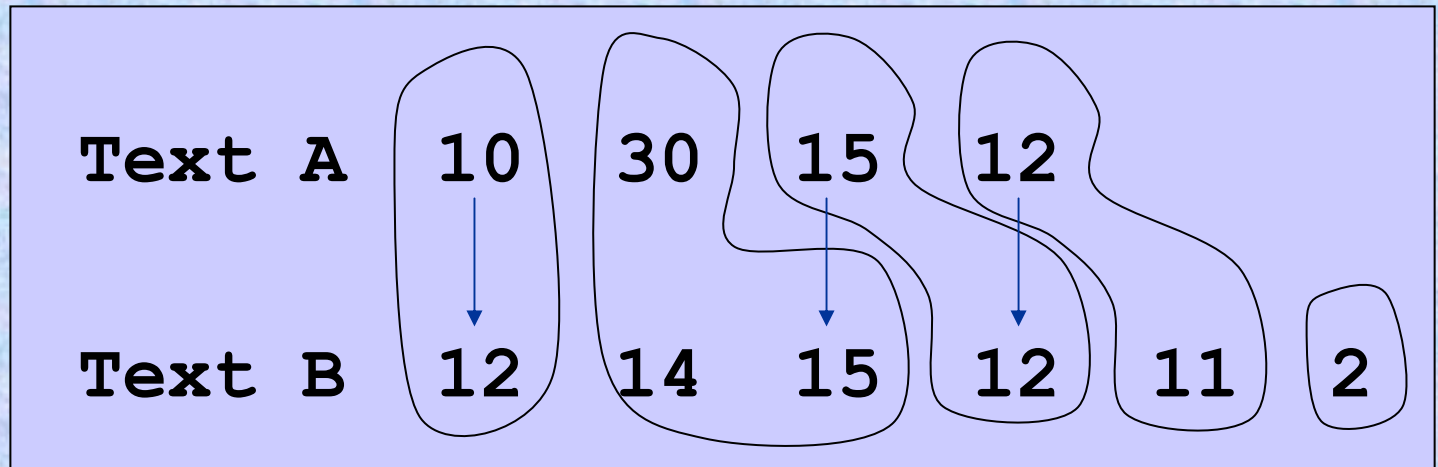
- Involving traditional analysis, e.g.
 - parsing
 - tagging
 - use of bilingual lexicon
- Fully automatic
 - simple heuristics (length-based)
 - identification of “anchors”
 - distributional statistics

- Some TM systems have built-in alignment tool

Length-based alignment

- Assumption of significant correlation in the relative length of sentences, measured in words or, better, characters.
- Major flaw: assumes that there is a natural common segment type (“sentence”) ... not true, e.g. for English and Chinese.
- Also depends on “closeness” of translation
- Can be improved by looking for “anchors”

Lengths of sentences in the two texts



Bead type: 1:1 1:2 1:1 1:1 0:1

“Cognates” as anchors

- “Cognates” not defined historically, but purely symbolically, i.e.
- Word pairs with a sufficient percentage of characters in common
 - Includes literals (names, dates, numbers) and bits of markup
- Applies in more limited way across writing systems ...
- Looking for transliterations may be fruitful, if transliteration scheme is straightforward
- A bilingual machine-readable dictionary (if available) can help

- Mark-up (especially HTML pages) can provide a good set of anchors
- Not always straightforward however.

```
<HTML>
<HEAD><TITLE>ATLAS
  Symposium</TITLE></HEAD>
<BODY bgcolor="ffffff" text="115511"
  LINK="004080" vLINK="0040800">
<center>
<h1>Arabic Translation and Localisation
  Symposium<p>Symposium sur la
  Traduction et la Localisation en
  Arabe<br></h1>
...
</center>
<p>It is one of the five official
  languages of the United Nations, it
  has 260 million native speakers, and
  is used as a second language by a
  further 1.3 billion people.
...
<center>
<li>Arabic corpus processing
<li>Development of Arabic resources
<li>Web tools for Arabic
```

```
<HTML>
<HEAD><TITLE>Symposium ATLAS</TITLE></HEAD>
<BODY TEXT="#115511" BGCOLOR="#FFFFFF"
  LINK="#004080"VLINK="#048000">
<CENTER><IMG SRC="..." ALT="logo"
  HEIGHT=145 WIDTH=184>
<H1>Symposium sur la Traduction et la
  Localisation en Arabe<P>Arabic
  Translation and Localisation
  Symposium<BR><IMG
  SRC="arabatlas.gif"></h1>
...
</CENTER>
<p>L'une des cinq langues officielles de
  l'ONU est l'<B>arabe</B>, la
  languematernelle de 260 millions de
  locuteurs, qu'utilisent environ 1.3
  milliards de musulmans comme
  deuxi&egrave;me langue.
...
<CENTER>
<LI>les standards de codage des
  caract&egrave;res arabes</LI>
<LI>le traitement des corpus en arabe</LI>
<LI>le d&eacute;veloppement des ressources
  pour l'arabe</LI>
<LI>les outils Internet pour l'arabe</LI>
```

Distribution-based word alignment

- Extraction of bilingual vocabulary
- May help the sentence-alignment problem, or be a side-effect of it
- Based on assumptions:
 - Words have one sense per corpus
 - Words have a single translation per corpus
 - There are no missing translations
 - The frequencies of words and their translations are comparable
 - The positions of words and their translations are comparable

Identifying likely word pairs

Contingency matrix:

Divide texts (i,j) into an equal number of equally sized segments

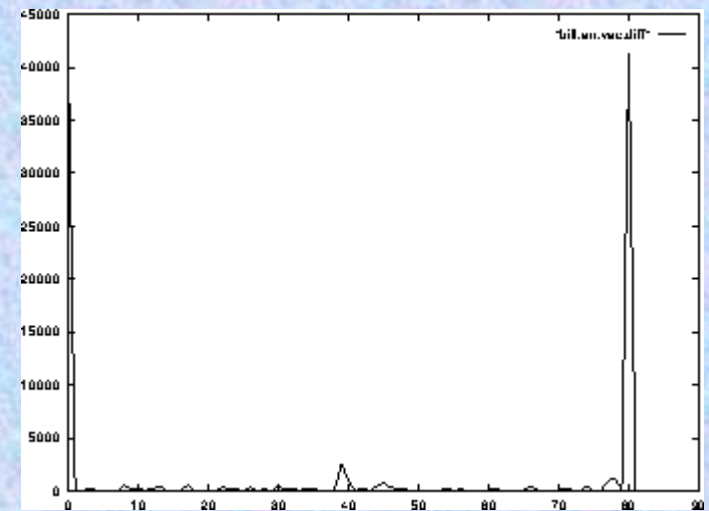
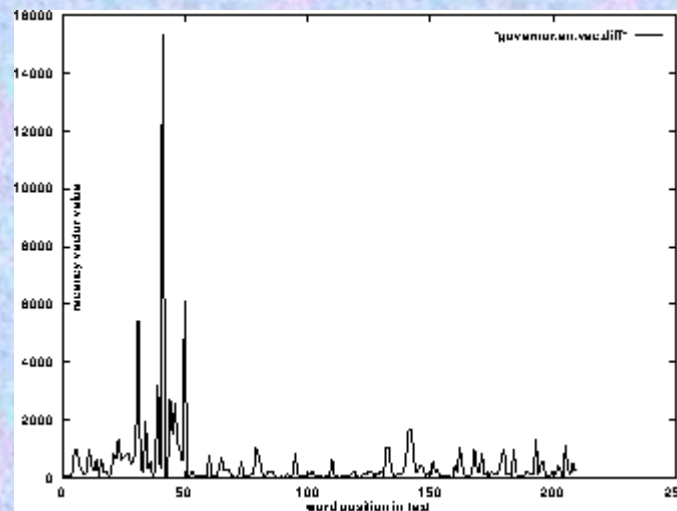
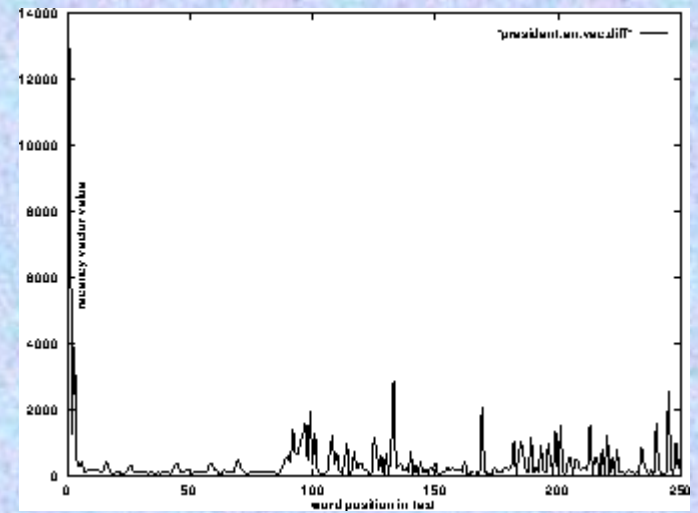
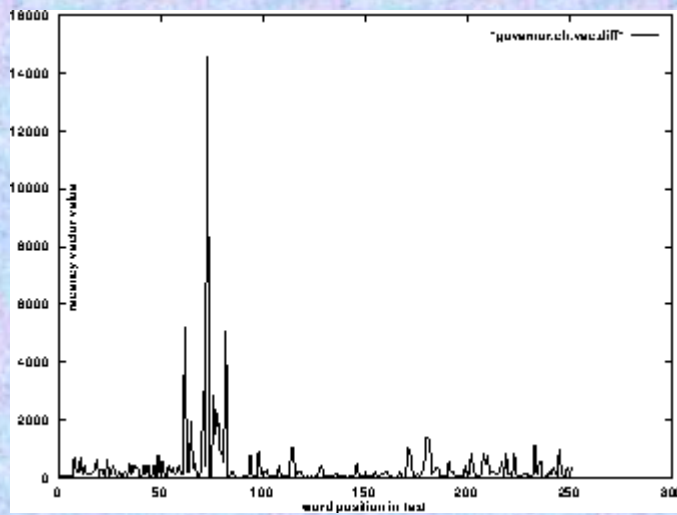
For a candidate word pair, plot total number of segments in which both occur (a), one but not the other (b,c) and neither (d)

Simple calculation using these four figures to determine whether distribution is significantly similar

	i	$\sim i$
j	a	b
$\sim j$	c	d

Recency vectors

- Measure the offset (in characters) between each occurrence of each word
- This gives a “recency vector”: a sequence of integers - each word has a characteristic trace
- Recency vectors for candidate word pairs can be compared using Levenshtein Distance



Chinese word for ‘governor’ compared to *govenor*, *president* and *bill*
 (Fung & McKeown 1994:84)

How well do they work?

- **Mixed results are reported**
- **Depends on language-pair, how “well-behaved” the corpus is, etc.**
- **Assumptions listed above are not all true**
 - **Words have one sense per corpus**
 - **Words have a single translation per corpus**
 - But capitalization and morphology must be normalized
 - **There are no missing translations**
 - Not always the case
 - **The frequencies of words and their translations are comparable**
 - Yes, but so are collocations
 - **The positions of words and their translations are comparable**
 - Only to a certain extent

Terminology and multiword equivalents

- Assumption of 1:1 word correspondence invalid
- Multi-word equivalents can be identified
- Look for monolingual collocations (using similar measures to those already seen)
- Look for high scoring combinations in bilingual correspondences

Translation Memory and EBMT

- TM lacks ability to make use of the example
- EBMT has two additional stages after matching: (segment) **alignment** and **recombination**
- Alignment identifies reusable portions of target-language segment associated with matched source-language phrase
- Recombination assembles the portions

Example (Sato & Nagao 1990)

Input

He buys a book on international politics

Matches

He buys a notebook.

Kare wa nōto o kau.

I read a book on international politics.

Watashi wa kokusai seiji nitsuite kakareta hon o yomu.

Result

Kare wa kokusai seiji nitsuite kakareta hon o kau.

Derivation of translation templates

I took a ticket from Mary \Leftrightarrow *Mary'den bir bilet aldım*
I took a pen from Mary \Leftrightarrow *Mary'den bir kalem aldım*

I took a ... from Mary \Leftrightarrow *Mary'den bir ... aldım*
ticket \Leftrightarrow *bilet*
pen \Leftrightarrow *kalem*

Exercise

A man saw the girl \Leftrightarrow *Mae dyn wedi weld y* ferch

A man saw the dog \Leftrightarrow *Mae dyn wedi weld y* ci

girl \Leftrightarrow *ferch*

dog \Leftrightarrow *ci*

A man saw the ... \Leftrightarrow *Mae dyn wedi weld y* ...

A teacher saw the book \Leftrightarrow *Mae athro wedi weld y llyfr*

{teacher, book} \Leftrightarrow {*athro, llyfr*}

A [j] saw the [k] \Leftrightarrow *Mae [j] wedi weld y [k]*

teacher \Leftrightarrow *athro*

book \Leftrightarrow *llyfr*

~~a \Leftrightarrow *mae*~~

~~the \Leftrightarrow *y*~~