On the Pleasures of being Bi-textual …

HLT-NAACL 2003 Workshop:

Building and Using Parallel Texts

Data-driven MT and Beyond

OR:
My life in parallel text

Elliott Macklovitch

Laboratoire RALI

Université de Montréal

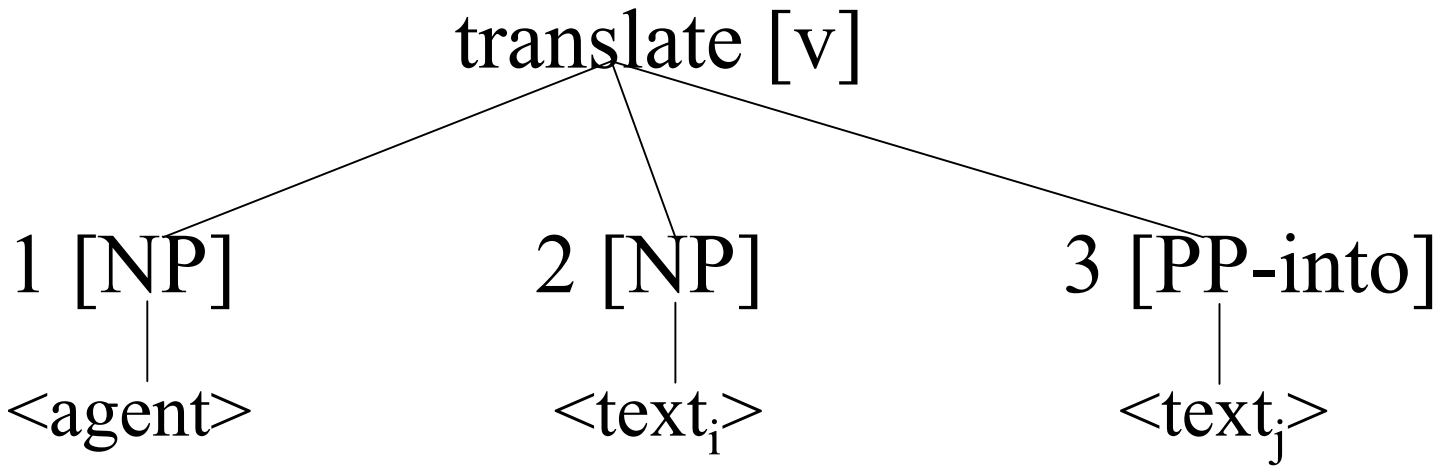
Acknowledgements

- I'm very flattered by this invitation…
  - but I'm not going to take it too personally
- The privilege of having worked with some remarkably talented researchers in NLP
  - acknowledge my debt to friends & colleagues
- A synopsis of RALI's work in parallel text
  - introduction that will hopefully "set the table" for more detailed presentations to follow

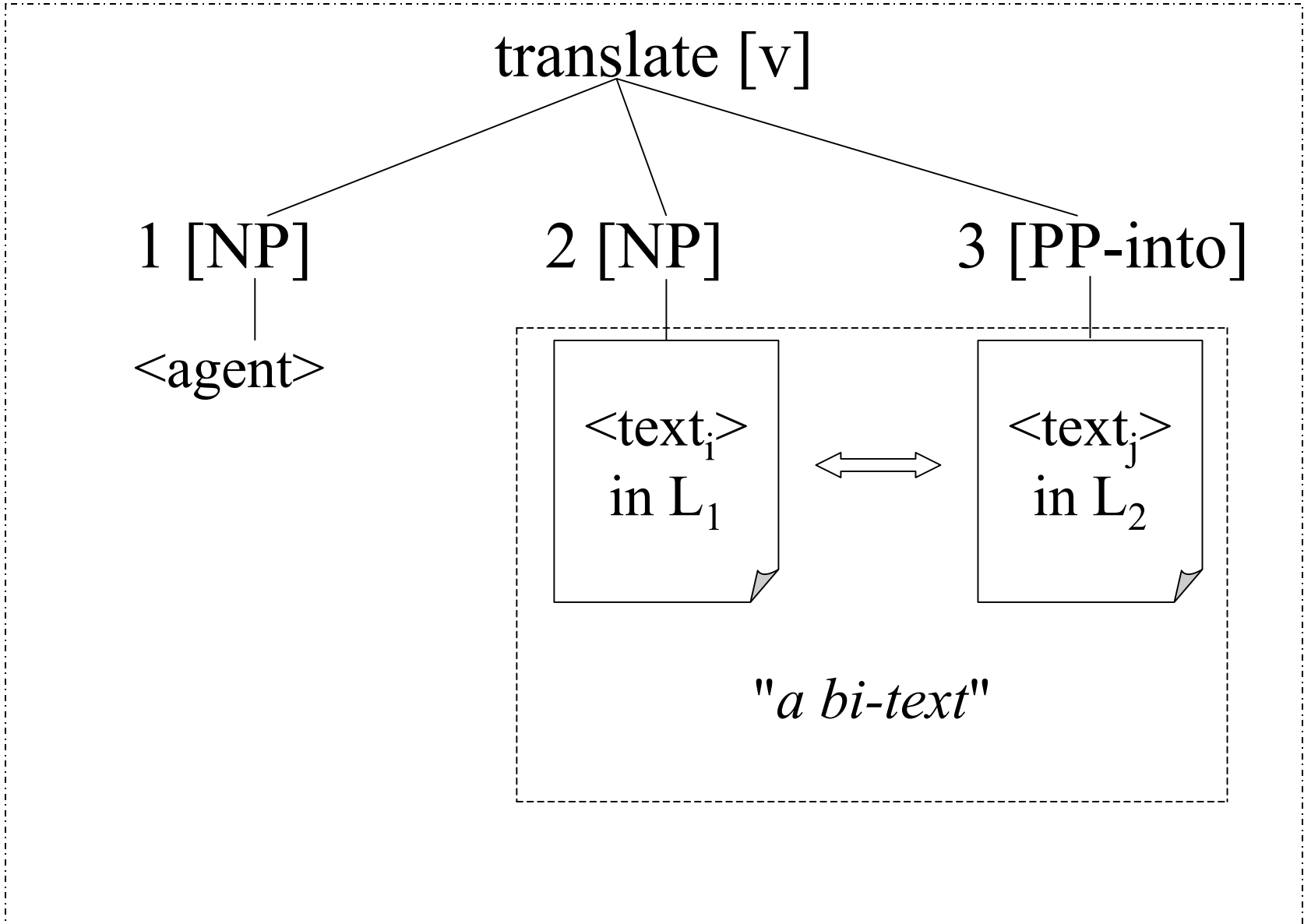
Early History

- (Melby, 1981):
  - 1st known proposal to store past translations electronically for bilingual concordancing
- (Harris, 1988a, 1988b):
  - coins the term "bi-text"
- (Gale & Church, 1991), (Brown et al, 1991)
  - 1st published algorithms for aligning sentences in parallel text

Definitions – (1)



- $text_i$ is a (pre-existing) source text
- TR's job is to produce target $text_j$, in a different L
- mean-preserving relation between $text_i$ & $text_j$


translate [v]
1 [NP]
2 [NP]
3 [PP-into]
<agent>
<text$_i$> in L_1
<text$_j$> in L_2
"*a bi-text*"

Definitions – (2)

$$\text{text}_i \leftrightarrow \text{text}_j$$
$$\text{text}_k \leftrightarrow \text{text}_l$$
$$\text{text}_m \leftrightarrow \text{text}_n$$
$$\ldots.$$

- a collection of bi-texts constitutes a *parallel corpus*

Definitions – (3)

- translation is a *transitive* relation
- given:

$$text_i \leftrightarrow text_j \leftrightarrow \ldots text_n$$

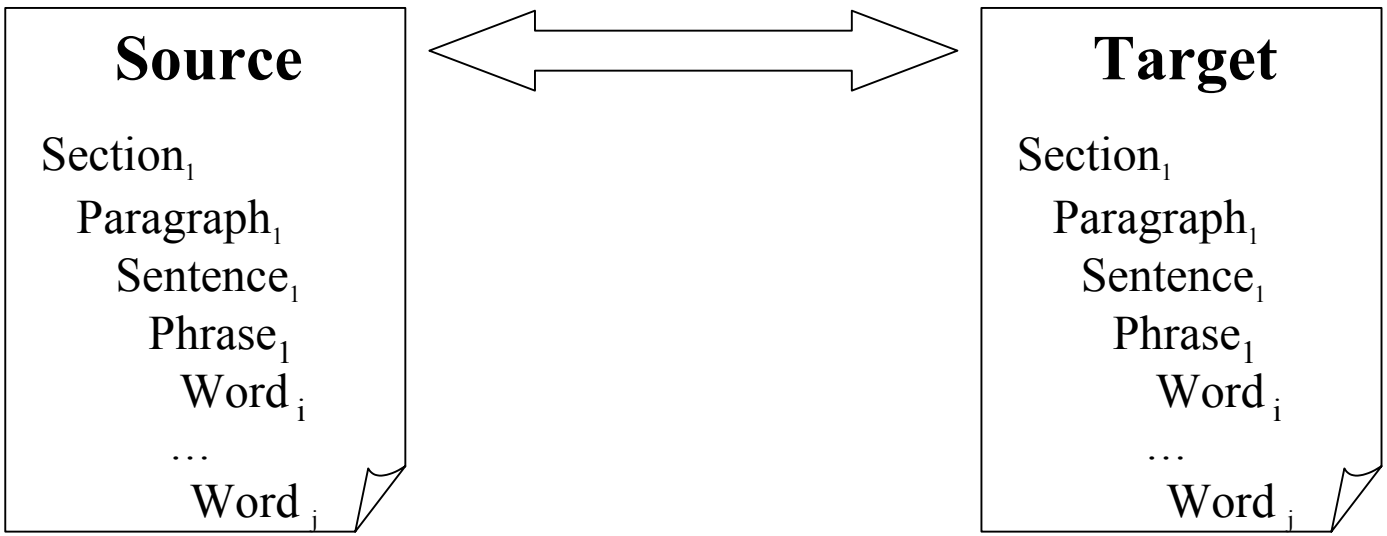
then $text_n$ is a translation of $text_i$

- the collection of $texts_{i\text{-}n}$ also constitutes a parallel corpus

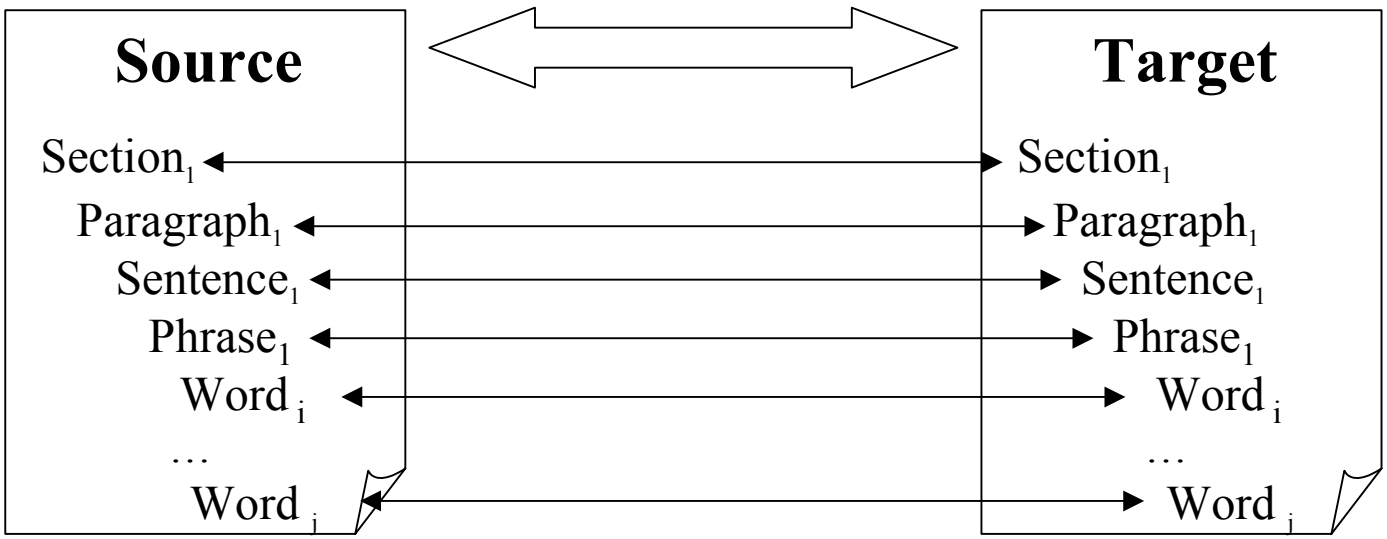
Translation is *compositional*

- the translation T of some textual segment S is a function of the translation of the sub-segments s_1, s_2,…s_3 that compose S
- compositionality can be applied recursively to two texts that are mutual translations, i.e. to progressively smaller textual units

Hierarchical correspondences



Hierarchical correspondences



Translation relation: $tr_{L1,L2}(S,T)$

- historically, efforts have focussed on the productive characterization of this relation
  - given S, define a procedure that will produce T
- can also be viewed as a recognition problem
  - given (S,T), decide if they are valid translations
- Translation Analysis aims to make explicit all the correspondences between S and T (Isabelle et al. 1993)

Definitions – (4)

"If we consider a text S and its translation T as two sets of segments S = $\{s_1, s_2, .., s_n\}$ and T = $\{t_1, t_2, ..., t_m\}$, an **alignment** A between S and T can be defined as a subset of the Cartesian product 2S X 2T, where 2S and 2T are respectively the set of all subsets of S and T. **The triple (S, T, A) will be called bi-text**." (Isabelle and Simard,1996)

Building Parallel Corpora

In the best of all possible worlds…

- large volumes of high-quality translation
  - freely available, in the public domain
  - ideally in well organized, parallel directories
  - with transparent naming conventions for parallel files
  - in format that allows for easy extraction of text
  - regularly updated
- = the Canadian Hansard!

Mining the Web for Parallel Texts

- *PT-Miner* (Chen & Nie, 2000)
  - search engines to locate candidate sites (specify an anchor to the other language)
  - host crawler to fetch max. no. of file names
  - file pairing algorithm generates possible names
  - apply various filters on downloaded files, e.g. file size, html structure, auto L-identifier, etc.
- used successfully to build STM for CLIR

Processing Parallel Text

- Extracting the text by deformatting
  - or do we exploit the formatting information to assist in the alignment?
- Segmenting the texts
  - a critical step!
  - difficult to properly align incorrectly segmented texts

Alignment

- The alignment *A* is intended to make explicit the correspondences between (S,T).
  - various levels of ***resolution***
- sentence alignment: largely solved
  - to the first length-based algorithms, (Simard, Foster & Isabelle, 1992) add dynamic cognates
  - (Véronis & Langlais 2000) for ARCADE results
  - "98.5% accuracy on 'normal' texts"

Word Alignment - 1

- A different kettle of fish!
- "bitext correspondence is typically only partial – many words in each text have no clear equivalent in the other text." (Melamed, 2000)

Word Alignment - 2

"Very often, it is difficult for a human to judge which words in a given target string correspond to which words in its source string. Especially problematic is the alignment of words within idiomatic expressions, free translations, and missing function words. … The problem is that the notion of correspondence between words is subjective." (Och and Ney, 2003)

Exploiting Parallel Corpora

MT and Translation Analysis

"In principle, translation analysis and MT are very similar problems. … But in cases where MT is not possible, we claim that it is still possible to build analyzers for the translations produced by human translators, and that there will be many uses for these devices." (P. Isabelle et al. 1993)

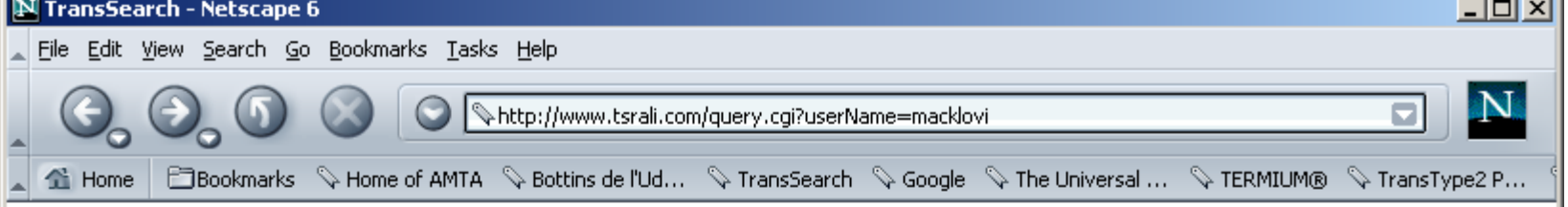
"The hierarchical model of translational correspondence implies a variable resolution parameter… [which] has no counterpart in MT (P. Isabelle, 1992)

Bi-textual Resolution

- low resolution bi-texts
  - representations that make explicit only a subset of all the correspondences between S and T
- TR production requires strong L-models
  - one cannot translate a paragraph without translating all its constituent elements
- in applying TR analysis to the development of translation support tools, one can often make do with weaker models

A new generation of translation support tools

"Existing translations contain more solutions to more translation problems than any other available resource." (P. Isabelle et al. 1993)



TransSearch

RALI

utilisateur: **macklovi**

Requêtes | Mon compte | Préférences | Aide | Quitter

Signet TransSearch
(qu'est-ce que c'est?)

Collection de documents : Hansard canadien (1986-2003)

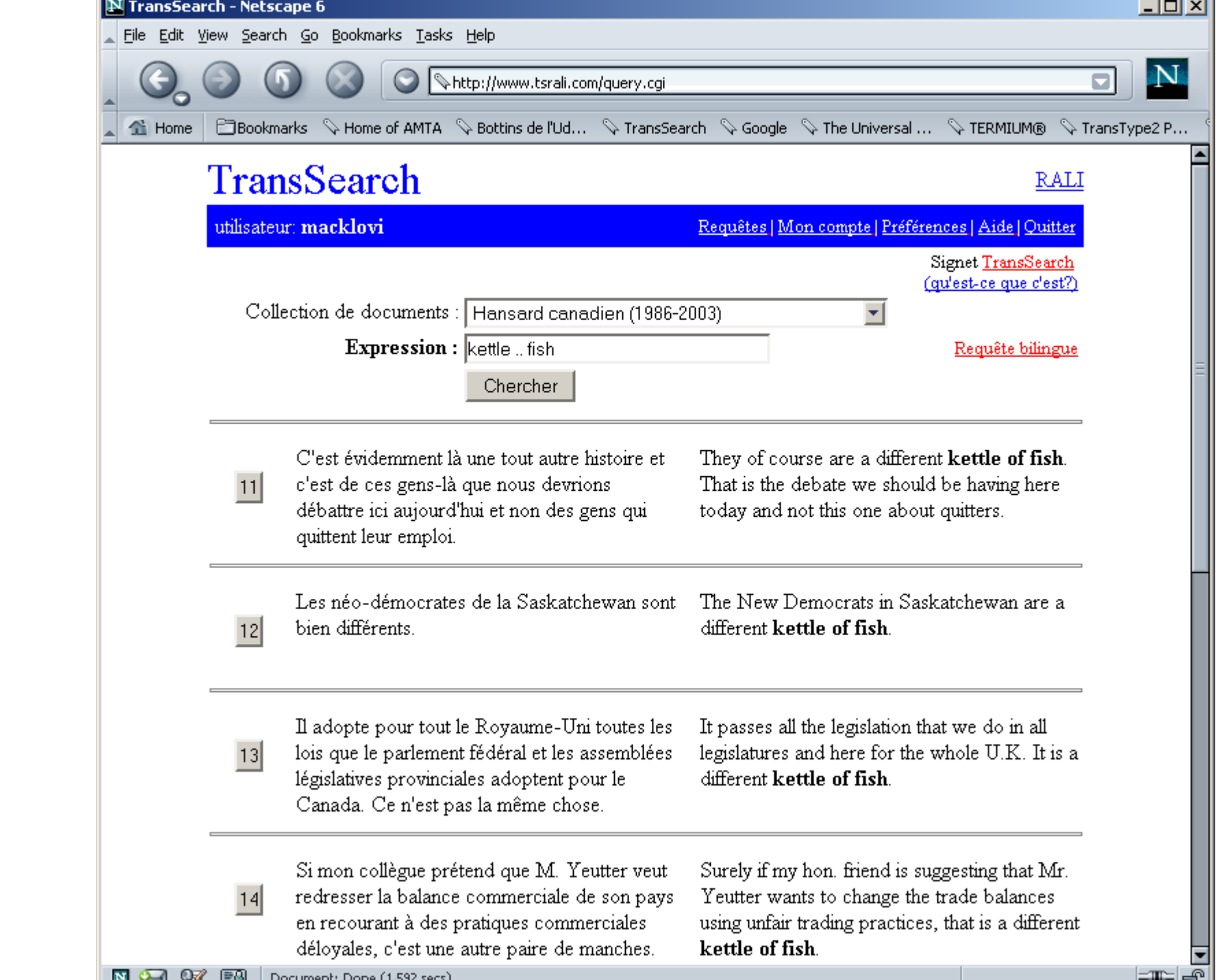
Expression : kettle .. fish

Requête bilingue

Chercher

Soumettez un mot ou une expression, en français ou en anglais : TransSearch cherchera des contextes où cette expression apparaît, de même que le contexte correspondant dans l'autre langue.

Webmestre


TransSearch - Netscape 6
File Edit View Search Go Bookmarks Tasks Help
http://www.tsrali.com/query.cgi
Home Bookmarks Home of AMTA Bottins de l'Ud... TransSearch Google The Universal ... TERMIUM® TransType2 P...
TransSearch
RALI
utilisateur: macklovi
Requêtes | Mon compte | Préférences | Aide | Quitter
Signet TransSearch
(qu'est-ce que c'est?)
Collection de documents : Hansard canadien (1986-2003)
Expression : kettle .. fish
Requête bilingue
Chercher
11
C'est évidemment là une tout autre histoire et c'est de ces gens-là que nous devrions débattre ici aujourd'hui et non des gens qui quittent leur emploi.
They of course are a different kettle of fish. That is the debate we should be having here today and not this one about quitters.
12
Les néo-démocrates de la Saskatchewan sont bien différents.
The New Democrats in Saskatchewan are a different kettle of fish.
13
Il adopte pour tout le Royaume-Uni toutes les lois que le parlement fédéral et les assemblées législatives provinciales adoptent pour le Canada. Ce n'est pas la même chose.
It passes all the legislation that we do in all legislatures and here for the whole U.K. It is a different kettle of fish.
14
Si mon collègue prétend que M. Yeutter veut redresser la balance commerciale de son pays en recourant à des pratiques commerciales déloyales, c'est une autre paire de manches.
Surely if my hon. friend is suggesting that Mr. Yeutter wants to change the trade balances using unfair trading practices, that is a different kettle of fish.
Document: Done (1.592 secs)

Le très hon. Brian Mulroney (premier ministre): Monsieur le Président, de toutes les questions que mon honorable collègue a posées depuis un certain temps, c'est certainement l'une des moins brillantes!	Right Hon. Brian Mulroney (Prime Minister): Mr. Speaker, this really, when we examine it, will be one of the least impressive questions asked by my hon. friend among a very long list that he has asked over a period of time.
J'ai dit au chef de l'opposition que je n'avais pas vu la déclaration de M. Yeutter. Je l'examinerai attentivement et je la comparerai avec l'interprétation du député de Winnipeg. Le 2 janvier, le président des États-Unis, qui est le patron de M. Yeutter a donné...	I told the Leader of the Opposition that I have not seen Mr. Yeutter's statement. I will look at it carefully and I will compare his statement with the interpretation placed upon that statement by the hon. gentleman from Winnipeg. The President of the United States, who is the boss of Mr. Yeutter, conveyed on January 2--
M. Boudria: C'est le vôtre aussi!	Mr. Boudria: He is your boss too.
M. Mulroney: Pardonnez-moi?	Mr. Mulroney: I am sorry?
M. Boudria: C'est aussi votre patron.	Mr. Boudria: He is your boss too.
M. Mulroney: ...l'impression que le gouvernement des États-Unis cherchait à intensifier ses échanges avec le Canada, avec le monde entier, tout comme nous. C'est exactement ce que nous faisons. Nous cherchons le moyen de libéraliser les échanges de façon à créer davantage d'emplois chez nous.	Mr. Mulroney: --the impression of the Government of the United States that it was seeking more trade with Canada, more trade around the world, which is our position. That is exactly what we are doing. We seek ways to liberalize trade so that we can create more jobs at home.
Si mon collègue prétend que M. Yeutter veut redresser la balance commerciale de son pays en recourant à des pratiques commerciales déloyales, c'est une autre paire de manches. Si oui, nous le traînerons très rapidement devant le GATT et devant nos autres commissions latérales. je citerai même peut-être le député de Winnipeg comme témoin de moralité. {section} L'immigration	Surely if my hon. friend is suggesting that Mr. Yeutter wants to change the trade balances using unfair trading practices, that is a different **kettle of fish**. If he said that, we will drag him before the GATT and before our binational panels very, very quickly. I might even bring in the Hon. Member from Winnipeg as a character witness. {section} IMMIGRATION
L'admission d'un terroriste reconnu--L'avertissement	Admission of convicted terrorist--Warning forwarded by

TSrali.com

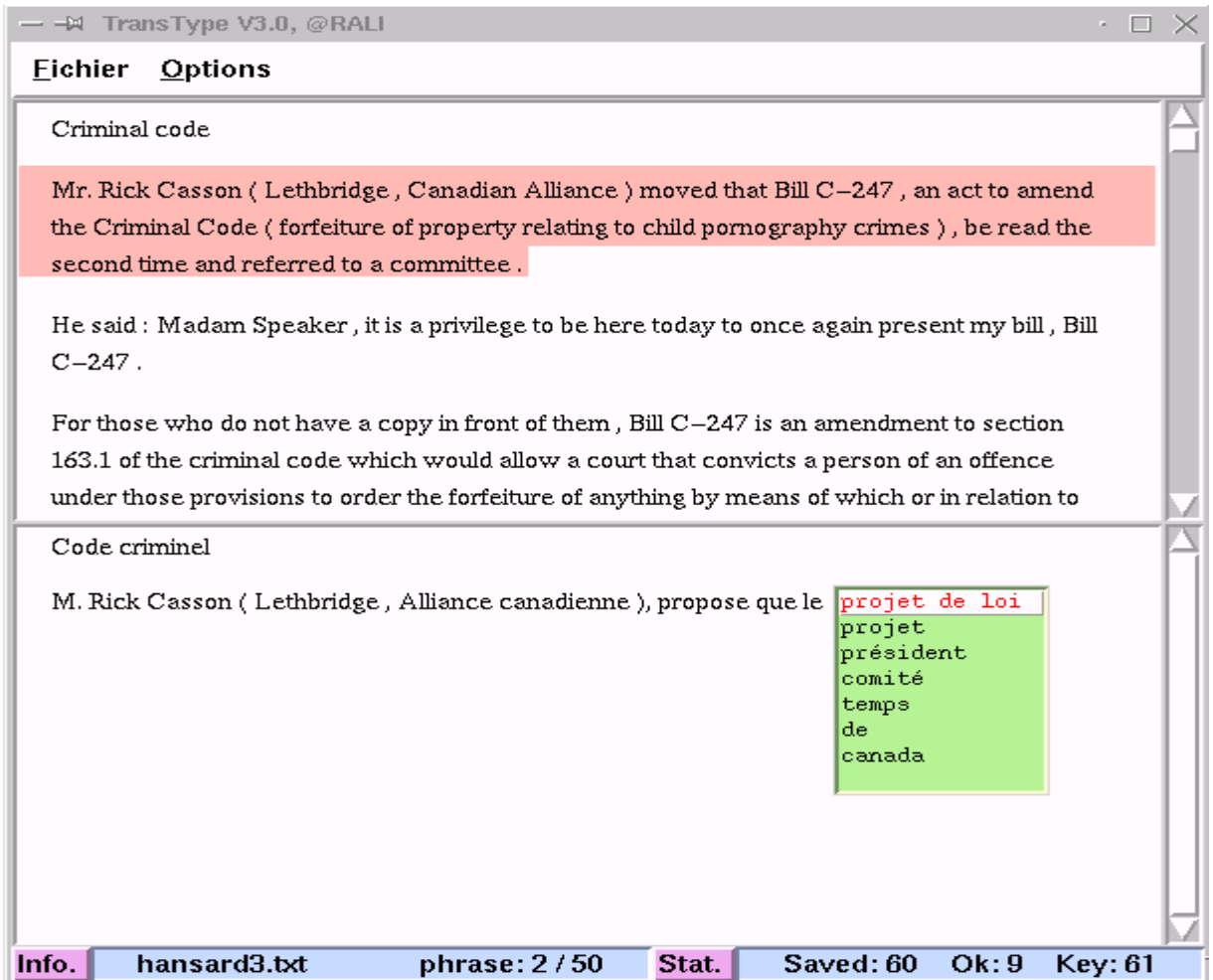
- Offered as an on-line subscription service
  - ~ 1500 subscribers; +75K queries per month
  - – Spanish-English DB to be added shortly
  - – Profitable enough to transfer to private sector
  - – HIGHLY APPRECIATED BY ITS USERS!

- System architect: Michel Simard

Beyond SMT?

- HQ translation is a moving target
  - there are often numerous good translations
  - even when an MT system manages to produce one, a human TR may well want to revise it
- *TransType*: a new approach to interactive MT
  - focus of the interaction is on the target text
  - TR in control; free to ignore system's proposals
  - completions ADAPT to changes in user input
  - for more details, see (Foster et al. 2002)

TransType: le prototype actuel



Other applications for parallel text

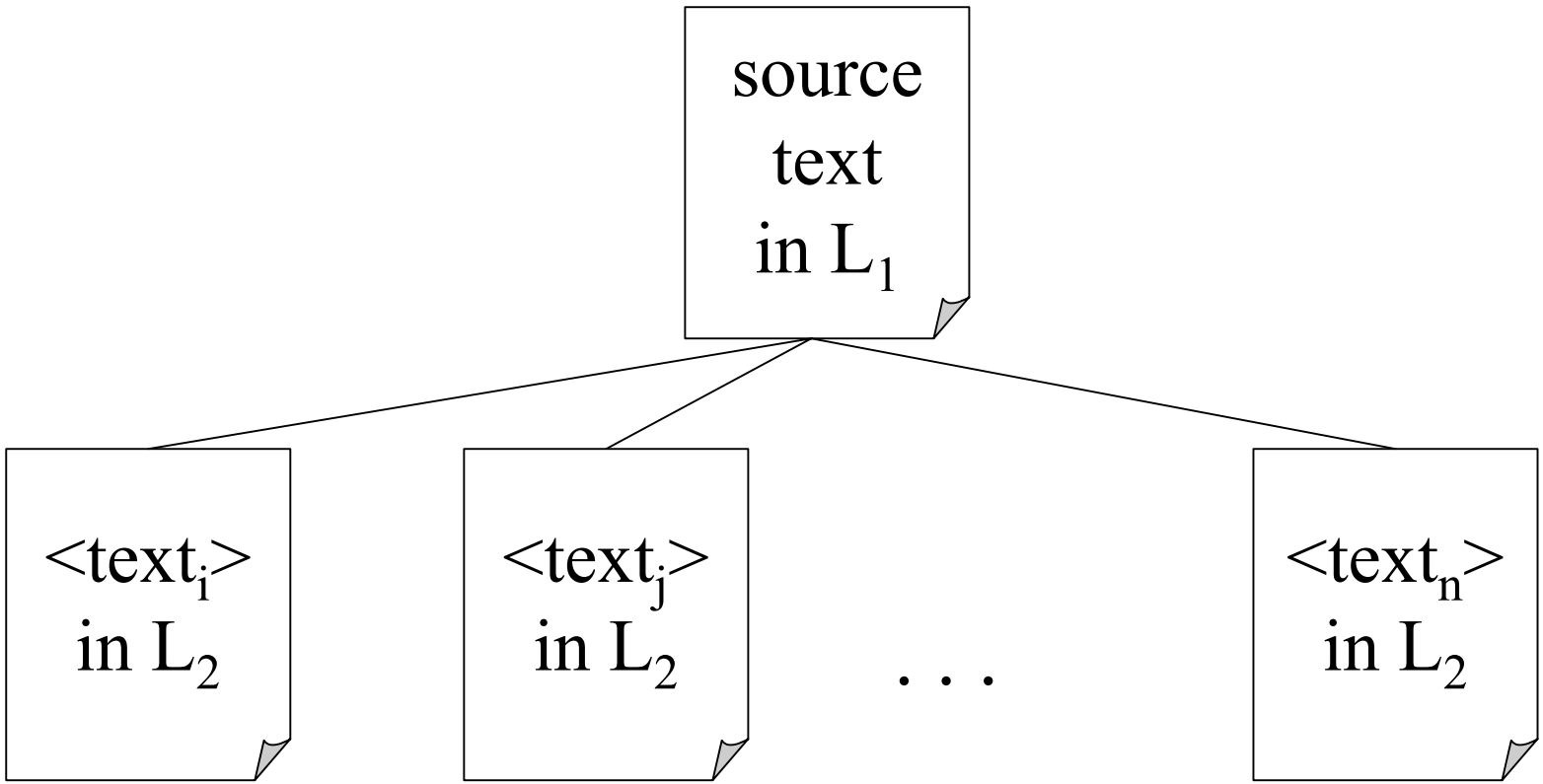
- Bilingual lexicon development
  - for human lexicographers, terminologists, etc.
  - methods for extracting from a parallel corpus the possible translations of each source word
  - doesn't provide for context-dependent selection
  - reliably identify non-compositional compounds and their translations
  - C.f. (Melamed 1998)

Word-sense disambiguation

"It would be a major breakthrough if the availability of parallel text made it possible to make progress on the sense disambiguation problem." …

"The fact that French and English are different as they are makes for a valuable research opportunity… We can use the French text to disambiguate word-senses in the English, producing a large sense-disambiguated corpus to develop and test word-sense disambiguation algorithms…"(Church & Gale 1991)

Multiple reference translations



Conclusion

- Parallel texts have certainly proven to be an fertile area for R&D in NLP
- I have attempted to "set the table" for the presentations that will follow in this WS
  - *Que la fête commence!*
  - *Let the festivities begin!*

References

Brown, Peter, J. Lai and Robert Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley CA, pp. 29-36.

Chen, J. and Jian-Yun Nie. 2000. Parallel Text Mining for Cross-language IR. In *Actes de la conférence RIAO*, Paris, pp. 62-77.

Church, Kenneth W. and William A. Gale. 1991. Concordances for Parallel Text. In Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, pp. 40-62.

Foster, George, Philippe Langlais and Guy Lapalme. 2002. User-friendly Text Prediction for Translators. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing,* Philadelphia PA.

Gale, William and Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley CA, pp. 177-183.

Harris, Brian. 1988a. Bi-text: A New Concept in Translation Theory. *Language Monthly*, no. 54, pp 8-10.

Harris, Brian. 1988b. Are You Bi-textual? *Language Technology*, no.7, p. 41.

Isabelle, Pierre. 1992. Bi-text: Toward a New Generation of Support Tools for Translation and Terminology. Published in French in *META*, 37(4), pp. 721-737.

Isabelle, Pierre, M. Dymetman, G. Foster, J-M. Jutras, E. Macklovitch, F. Perrault, X. Ren and M. Simard. 1993. Translation Analysis and Translation Automation. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan, pp. 12-20.

Isabelle, Pierre and Michel Simard. 1996. Propositions pour la représentation et l'évaluation des alignements et des textes parallèles. Rapport technique du CITI. Laval (QC), Canada. (http://www-rali.iro.umontreal.ca/arc-a2/PropEval)

Melamed, I. Dan. 1998. Empirical Methods for MT Lexicon Development. In *Proceedings of the Third Conference for Machine Translation in the Americas, AMTA'98*, Langhorne PA, Springer-Verlag, LNAI 1529, pp. 18-30.

Melamed, I. Dan. 2000. Models of Translational Equivalence among Words. *Computational Linguistics*, 26(2), pp. 221-249.

Melby, Alan. 1981. A Bilingual Concordance System and its Use in Linguistic Studies. In *Proceedings of the 8th Lacus Forum*, Hornbeam Press, Columbia SC, pp.541-54.

Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): pp.19-51.

Simard, Michel, George Foster and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada, pp. 67-81.

Véronis, Jean and Philippe Langlais. 2000. Evaluation of parallel text alignment systems : The Arcade project. In *Parallel Text Processing*, ed. Jean Véronis, Kluwer Academic Publishers, pp. 369-388.