Université de Montréal

Recent developments in the *TransSearch* bilingual concordancer

Guy Lapalme
RALI - Université de Montréal
joint work with
Fabrizio Gotti, Philippe Langlais, Julien Bourdaillet, Stéphane Huet, Jacques Steinlin
Industrial contributor: Terminotix

rali

Recherche appliquée en linguistique informatique

Applied Research in Computational Linguistics

RALI's personnel includes computer scientists and linguists with considerable experience in natural language processing. It is the largest university NLP laboratory in Canada.

Research Projects

- Translation
- Information extraction
- Automatic summarization
- Information retrieval
- Judicial texts processing
- Environmental information

System demos

- Translation
  - TransSearch : bilingual concordancer
  - TransType : animation of an interactive translation session
- Reacc : automatic French accentuation
- SILC : language and coding detection
- Lexiqum : Québec text concordancer

Information

- Publications
- Textual Resources

- Contacts
- Members
- Collaborators

- RALI_OLST-MITACS seminars (mostly in French)

- Courses by professors of the RALI
  - Fall 2011: IFT6810: Traitement Statistique des Langues Naturelles
  - Fall 2011: IFT3335: Intelligence Artificielle

- *RALI in the press*

A selection of international conferences in NLP

RALI - Today

- 3 professors
  - Philippe Langlais : machine translation
  - Jian-Yun Nie : information retrieval
  - Guy Lapalme : summarization, generation, info extraction
- Adjunct professor : Atefeh Farzindar
- Students
  - 1 post-doc
  - 8 Ph.D.
  - 5 M.Sc.
- 1 Research associate
  - Fabrizio Gotti

rali

Goals of RALI

- Applied NLP using symbolic and statistical methods
  - Tools for translators
  - Writing aids
- Collaborative work
  - Linguistics department
  - Industrial partners
  - Governmental partners

rali

http://www.tsrali.com

TRANSSEARCH TERMINOTIX rali

USER: lapalme | QUERIES | MY ACCOUNT | PREFERENCES | CONTACT | HELP | QUIT

Personalized Favorite / Bookmark : TransSearch (what is this?) Bilingual query

Document collection : House of Commons Hansard (1986-2011)

Expression: take+ .. ride Search

1	Ils peuvent se faire passer des « p'tites vites ».	They can be **taken for a ride**.
2	Voyez ce qui leur arrive lorsqu'on profite d'elles.	Look what happens when they are **taken for a ride**.
3	C'est trop embêtant et on ne soupçonne pas les compagnies de vouloir nous embobiner.	It becomes too cumbersome and we think nobody is **taking us for a ride**.
4	On le prend pour aller se promener et on le laisse dans un endroit ou dans un autre.	They **take it for a ride** and leave it somewhere else.
5	Ils ont été abordés par des gens qu'ils considéraient comme des amis, en qui ils avaient confiance, et ils se sont fait rouler.	They had been approached by people who they considered friends and trusted and they were **taken for a ride**.
6	Il serait donc important de connaître toute l'information sur les revenus, si on ne veut pas être les dindons de la farce.	That is why it is important to have full access to information about revenue if we do not want to be **taken for a ride**.

rali

Timeline of TransSearch

1992 Michel Simard (CITI) on Sun machines
1995 First Web version (free)
1997 Web at RALI
1999 Paying subscribers - TSRali.com
2003 Transfer to Terminotix
2008 NSERC - CRD project for new version
2011 TransSearch with *translation spotting* in beta

Julien Bourdaillet, Stéphane Huet, Philippe Langlais and Guy Lapalme. **TransSearch: from a Bilingual Concordancer to a Translation Finder.** *Machine Translation*, vol. 24, number. 3-4, p. 241-271, Dec 2010

rali

TRANSSEARCH 3 BETA TERMINOTIX rali

UTILISATEUR : **lapalme** | REQUÊTES | MON COMPTE | PRÉFÉRENCES | AIDE | QUITTER

Signet / Favori personnalisé : TransSearch (qu'est-ce que c'est ?) | Requête bilingue

Collection de documents : Les Hansards canadiens

Expression : take+ .. ride | Chercher

92 traductions de ***take+ .. ride*** dans 106 occurrences

dindons de la farce	4
monté un bateau	3
faire avoir	3
se fasse rouler	2
fait berner	2
se fait jouer	2
moqués de	2
fait	2
les a	2
se sont fait avoir	2
le public pour attirer la	1
a fait une ballade	1
nous rouler dans ce projet nous tous	1
en train de monter un bateau à la population canadienne	1
tête des contribuables que se paie le	1
passer une petite vite	1
bourrer de l'autre côté de la chambre en	1
ont pris la voiture que pour faire une balade	1

dindons de la farce 4

Emissions continue to rise and taxpayers are being **taken along for the ride**.	Les émissions continuent d'augmenter et c'est le contribuable qui est **le dindon de la farce**.
They are left with nothing. Now they are here illegally with no documentation. Canadians are being **taken for a ride**.	Ces personnes se trouvent ici illégalement, elles n'ont aucun document et nous, les Canadiens, sommes les **dindons de la farce**.
This would affect close to 400,000 Canadians, 80,000 of them Quebecers, who have been the ones **taken for a ride**.	Il s'agit d'une mesure qui toucherait près de 400 000 Canadiens, dont 80 000 Québécois, qui ont été les **dindons de la farce**.
I think that this is a prime example of a tainted system in which people who cannot afford to invest in sectors eligible for tax credits are urged to do so through all kinds of scams and end up being **taken for a ride**.	Je pense que c'est un exemple patent d'un système vicié, où des gens qui n'ont pas les moyens d'investir dans des domaines où on peut obtenir des crédits d'impôt se voient, par toutes sortes de subterfuges, invités à le faire et, en bout de ligne, ils se trouvent à être **les dindons de la farce**.

rali

TRANSSEARCH 3 BETA TERMINOTIX rali

UTILISATEUR : **lapalme** REQUÊTES | MON COMPTE | PRÉFÉRENCES | AIDE | QUITTER

Signet / Favori personnalisé : TransSearch (qu'est-ce que c'est ?) Requête bilingue

Collection de documents : Les Hansards canadiens

Expression : take+ .. ride Chercher

92 traductions de ***take+ .. ride*** dans 106 occurrences

dindons de la farce	4
monté un bateau	3
faire avoir	3
se fasse rouler	2
fait berner	2
se fait jouer	2
moqués de	2
fait	2
les a	2
se sont fait avoir	2
le public pour attirer la	1
a fait une ballade	1
nous rouler dans ce projet nous tous	1
en train de monter un bateau à la population canadienne	1
tête des contribuables que se paie le	1
passer une petite vite	1
bourrer de l'autre côté de la chambre en	1
ont pris la voiture que pour faire une balade	1

monté un bateau 3

Mr. Yvan Loubier (Saint-Hyacinthe-Bagot, BQ): Mr. Speaker, the Minister of Finance is **taking us for a ride**.

M. Yvan Loubier (Saint-Hyacinthe-Bagot, BQ): Monsieur le Président, le ministre des Finances **nous monte un bateau**.

Ordinary Canadians are being **taken for a ride** by the same two old line parties which represent the corporate interest, the bottom line of this country, no matter what is going on.

Les Canadiens ordinaires se font **monter un bateau** par les deux vieux partis qui représentent les grandes sociétés qui ont toujours gain de cause, quelles que soient les circonstances.

It concerns the Meech Lake Accord. This week the Canadian Ethnocultural Council, the national umbrella organization representing multicultural associations across the country, had the following to say about the Accord. "Our worst suspicions about the process of the hearings were confirmed as the report merely rubberstamped the Accord...while we were **taken for a ride**...we cannot sit back and watch multiculturalism being set back for years".

Le Conseil ethnoculturel du Canada, l'organisation-cadre nationale représentant des associations multiculturelles de tout le pays, a réagi cette semaine en ces termes au rapport du comité chargé d'étudier l'Entente: «Nos pires soupçons au sujet des audiences se sont trouvés confirmés étant donné que le rapport n'a fait qu'entériner l'Entente... Ou nous a **monté un bateau**... Nous ne pouvons voir sans protester le multiculturalisme reculer de plusieurs années.»

rali

TransBases (1986-2011)

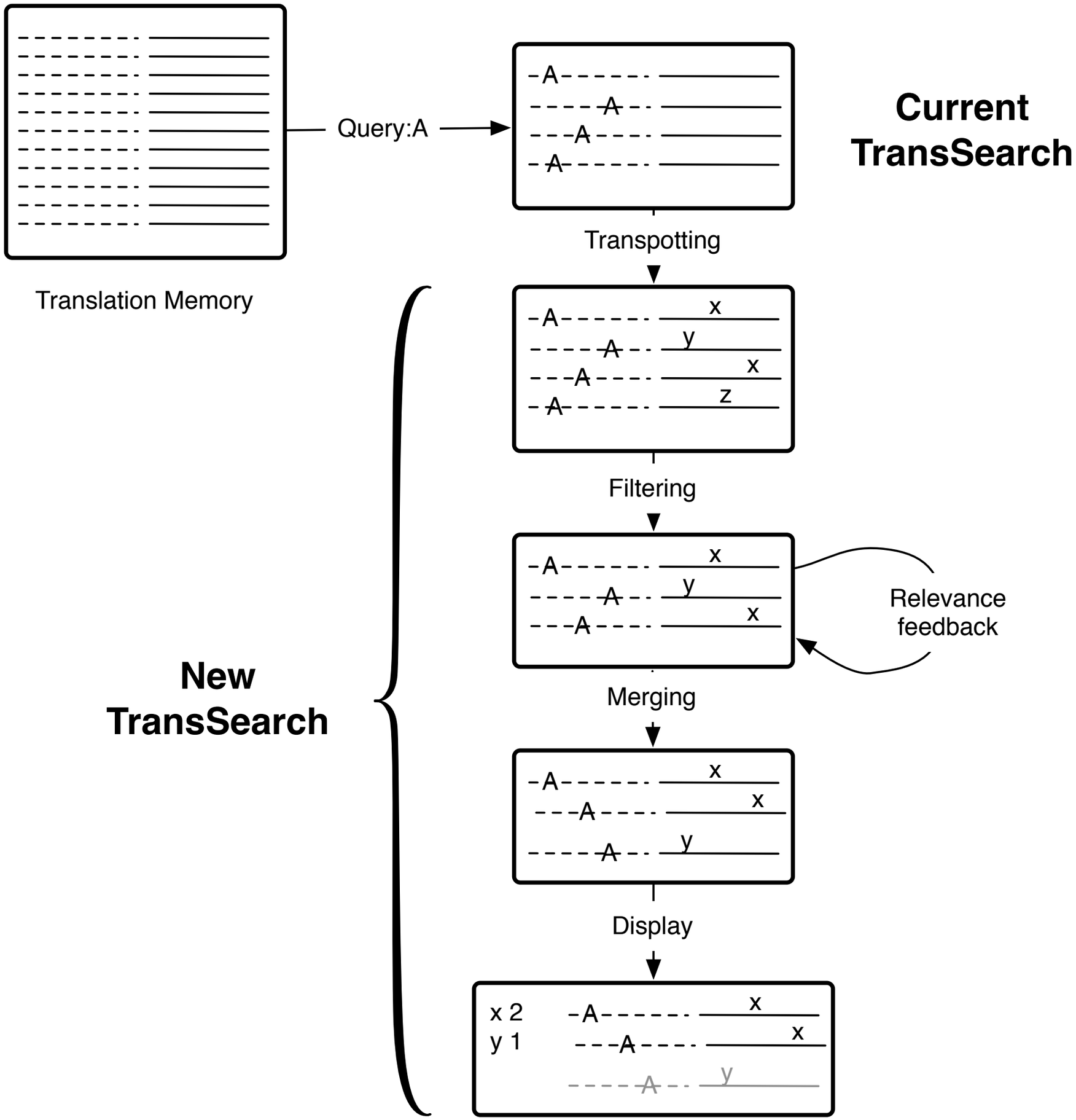
Corpus	docs	K pairs	M words
Hansard	2 854	10 500,0	339,6
Senate	1 025	1 200,0	47,6
Canadian courts	13 068	3 200,0	142,6
Meteorological alerts	2 398	7,2	0,2
	19 345	14 907,2	530,0

rali

TransBase Creation

- Finding texts and their translation
- Cleaning and normalization
- Sentence alignment
- Indexing

- Web based querying



Current TransSearch

New TransSearch

Translation Memory

Query:A

Transpotting

Current TransSearch

Filtering

New TransSearch

Relevance feedback

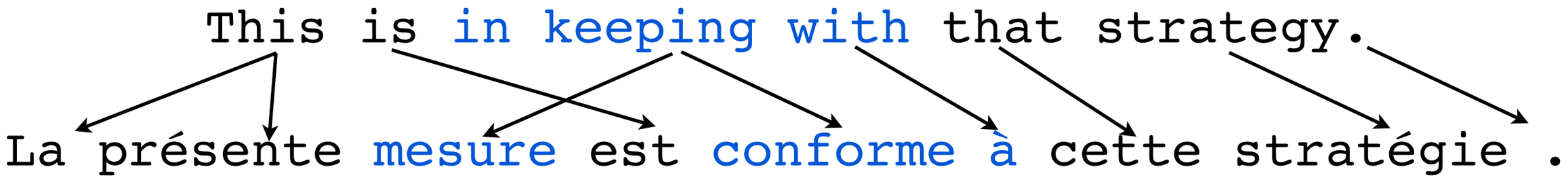
Merging

Display

x 2
y 1

rali

Transpotting



- Word alignment combining HMM and IBM models
  - combine alignments from both directions
  - add a contiguity constraint
- Phrase-based transpotting
  - use the phrase table produced by MOSES

rali

Postprocessing

on behalf of

Transpot	Frequency
au nom de	1424
au nom du	763
au nom des	683
de*	136
…	
dans l'intérêt des	15
de la part de	13
dans*	13
part de	13
…	
pour l'ensemble de	1
parler pour	1
loin que*	1
le bien*	1

Postprocessing

- Must deal with transpotting errors
  - between 20% and 40%
- User will focus on the first ten transpots
  - as many correct
  - as many diversified

rali

Postprocessing

- Filtering bad transpots using supervised classification
- Merging variants
  - inflectional form of canonical words
  - ignore some *grammatical* words
  - start merging by considering the most frequent ones
- Pseudo-relevance feedback
  - search for more frequent transpots in sentences with less frequent ones
  - develop a new transfer table from this specialized corpus and find new alignments

rali

Translation Memory

Query:A

Transpotting

Current TransSearch

New TransSearch

Filtering

Relevance feedback

Merging

Display

x 2
y 1

rali

Evaluation corpora

- Hansard (8.3 million Fr-En sentence pairs)
- 5000 most frequent English queries to TransSearch between 2001 and 2007
- 5000 sentence pairs for each query
- Bilingual phrase lexicon for testing
  - 2 358 entries (284 DEV, 2074 TEST)
  - average number of translations (3.6 DEV, 3.9 TEST)

rali

Evaluation process

- Transpotting
  - only the ones with the source containing the query and the target containing a translation from the dictionary (180 000 DEV, 1.4M TEST)
- Translation
  - comparison with bilingual phrase lexicon
- Classification
  - hand annotation of 531 queries (12 144 examples)

rali

Results

- Two best methods: `PBM` and `C-HMM-bi`
- Benefits from filtering and merging
- Limited impact of pseudo-relevance feedback

- But in the production version of TS3
  - at most 900 results for each query
  - transpotting
    - IBM-bi
  - filtering of transpots
    - 75% of grammatical words
    - only auxiliairies or numbers

rali

Non web uses of TransSearch

- Web service
  - application embedding
  - call from a text processor (e.g. Word)
- Educational use
- Integration *à la* TransType

Other bilingual concordancers

- With transpotting
  - TotalRecall (Wu, 2003)
    - Chinese-English word alignment
  - LinearB (Callison-Burch, 2005)
    - basic translation spotting
  - Linguee.com
    - many language pairs of web text (free with ads)
- Without transpotting
  - Opus
    - multiple parallel corpora
  - WeBiText (CNRC/Terminotix)
    - meta crawler for bilingual texts

rali

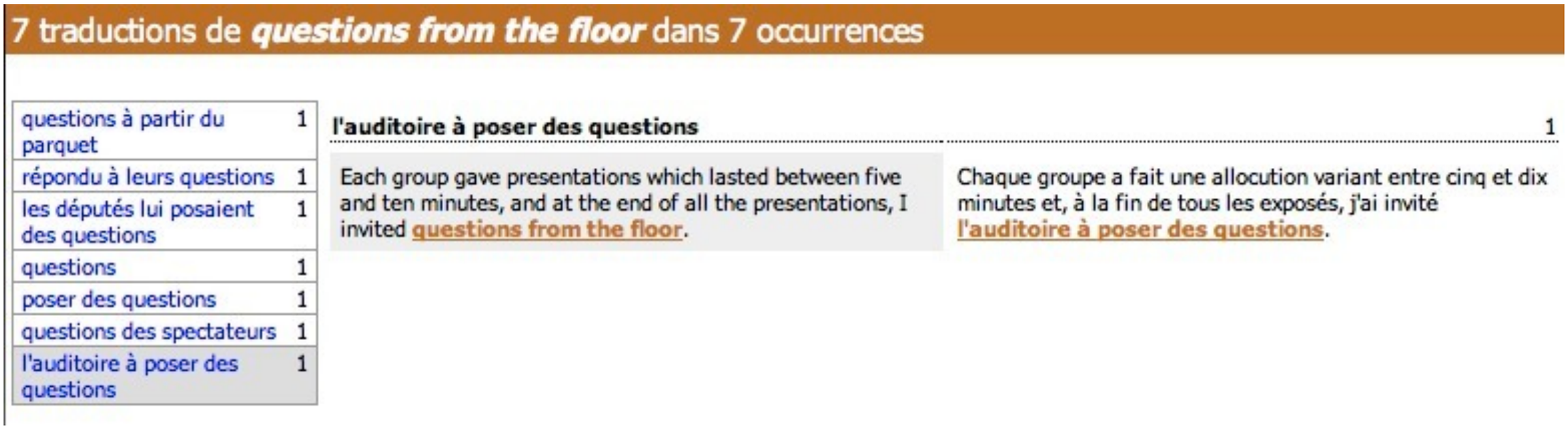
Why is TransSearch so popular ?

- Complements a terminology bank
- Not a bilingual dictionary

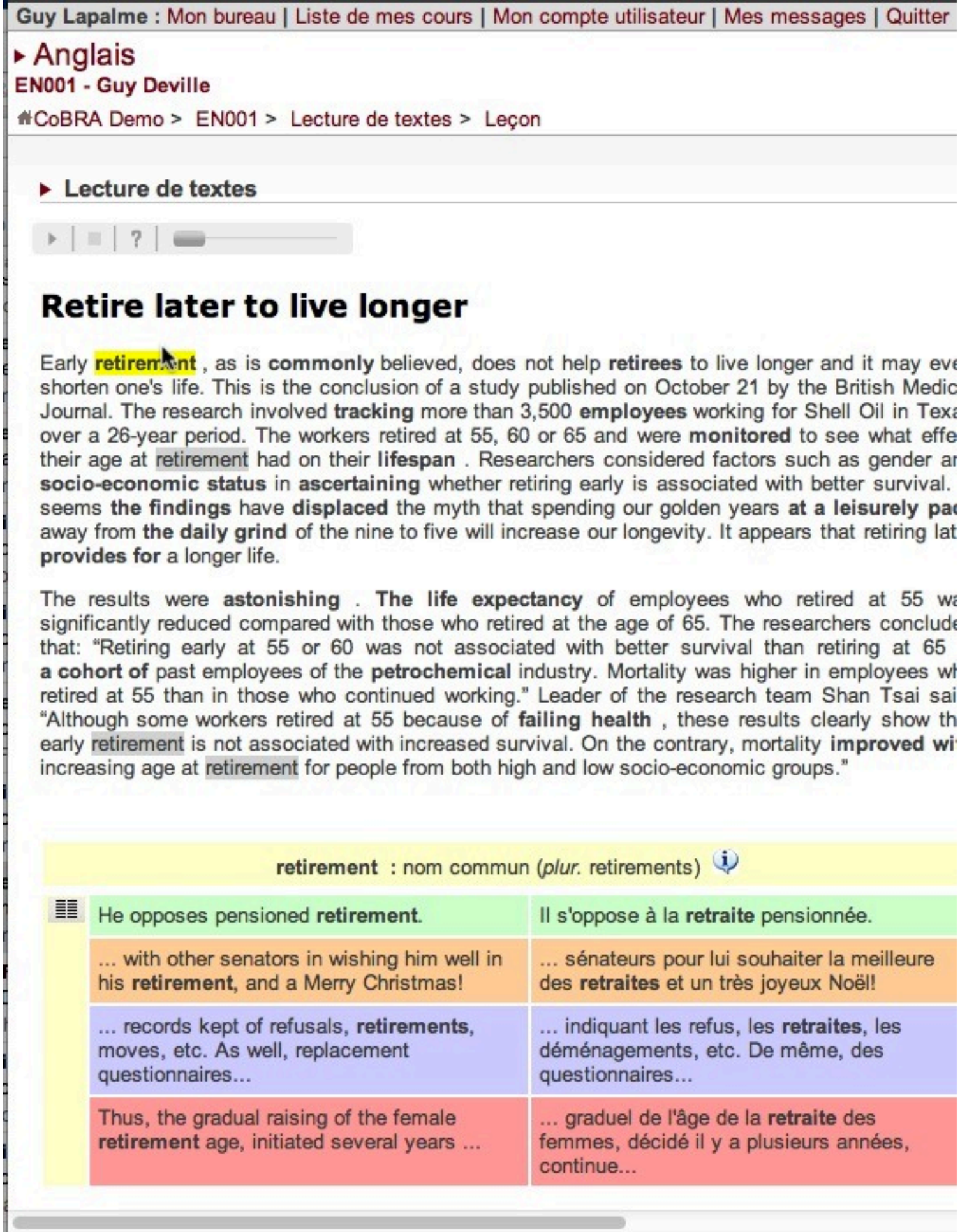
  87% of queries have more than one word
- Ready-made solution to translation problems
  - idiomatic expressions
  - non compositional translations
  - non literal translations

rali

Questions ?



rali

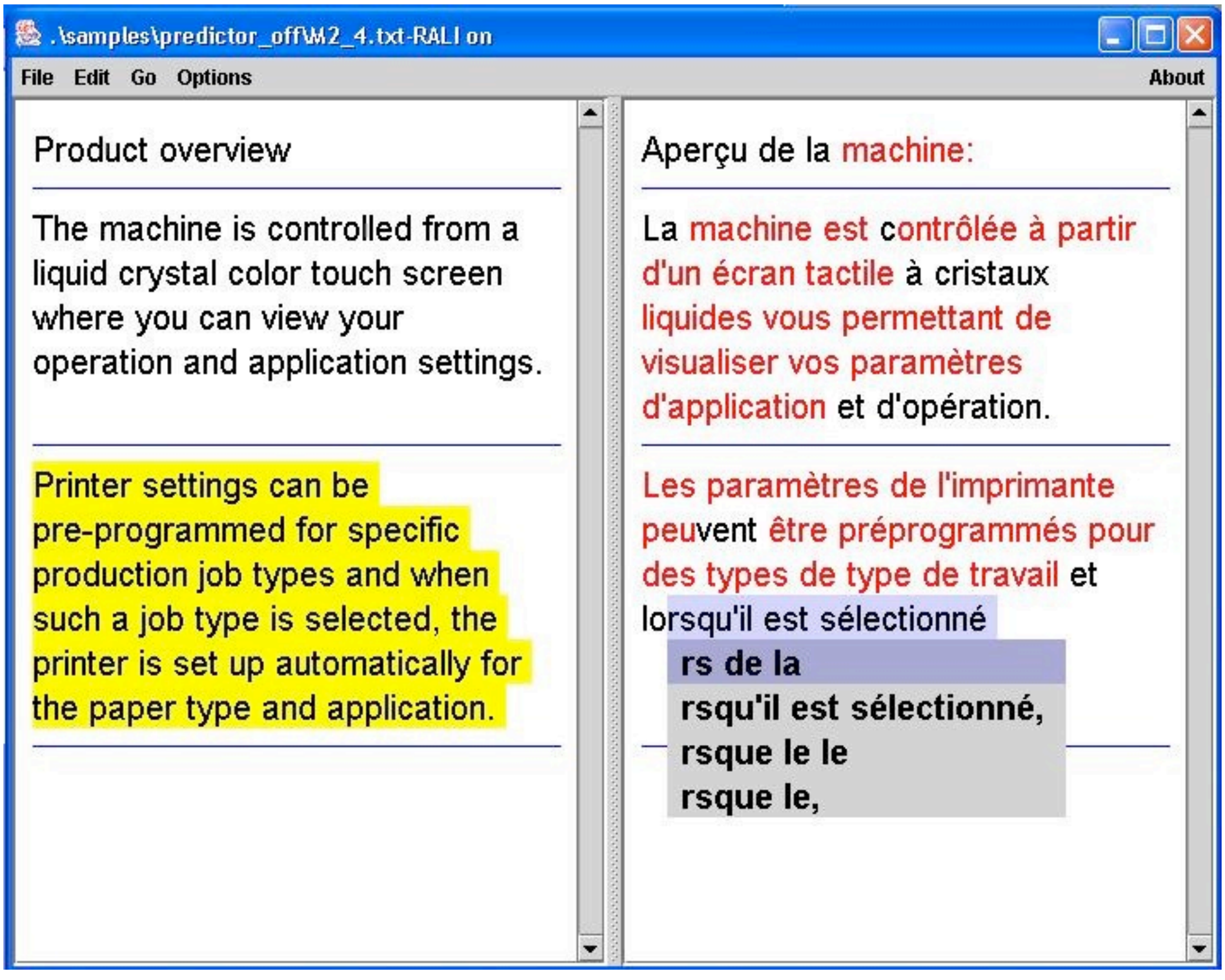


E-Lex

Deville et al. Université de Namur

rali

TransType



rali