Université de Montréal

University and Industry Partnership in NLP, is it worth the "trouble"?

Guy Lapalme
Université de Montréal

rali

This talk

- **is**
  - a personal account of experience
  - limited to Natural Language Processing grants related industry in Université de Montréal
  - a team effort
  - vague on the amounts of money
- **is not**
  - big (or even small) science
  - applicable in every case

rali

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 2009: IFT3335: Intelligence Artificielle
  - Fall 2009: IFT6810: Traitement Statistique des Langues Naturelles

- *RALI in the press*

A selection of international conferences in NLP

NLP at Université de Montréal

- TAUM (1970-1980)
- Incognito (Informatique cognitive) (1984-1997)
  - Expert Systems
  - Information Retrieval
  - NLP projects
    - Spelling checkers
    - Dictionary editing of Meaning-Text Theory
    - Deterministic parsers
    - Text generation

rali

Birth of the RALI

- Centre for Information Technologies Innovation (CITI) at Laval
  - Federal Government budget cuts in 1997
  - *Almost* privatized
  - Two teams were saved
    - Machine-Translation group (Pierre Isabelle *et al.*)
    - Computer tools for the handicapped

rali

RALI - Today

- 3 professors
  - Philippe Langlais : machine translation
  - Jian-Yun Nie : information retrieval
  - Guy Lapalme : summarization, IE, etc...
- Adjunct professor : Atefeh Farzindar
- Students
  - 4 post-docs
  - 7 Ph.D.
  - 5 M.Sc.
- 1 Research associate
  - Fabrizio Gotti

rali

Industrial collaboration

Expected benefits for industry

- Access to a rich talent pool
- Access to specific expertise
- Cost sharing using partnership governmental programs

rali

Industrial collaboration
Expected benefits for universities

- Dealing with realistic problems
- Inspiration for new research
- Motivation for students
- Alternative funding opportunities

rali

Types of industrial contributions

- Contracts (university overhead: 40%)
  - short term
- Licenses (uo: 40%)
  - existing technologies
- Grants (uo: 15%)
  - longer term, students involvement
- Partnership program (uo: 15% of indus. contrib.)
  - Governmental agencies (NSERC, Precarn, MITACS, Industry Canada) match an industrial contribution
- *Networks of Centres of Excellence / CFI*

rali

Types of industrial contributions

- Contracts (university overhead: 40%)
  - short term
- Licenses (uo: 40%)
  - existing technologies
- Grants (uo: 15%)
  - longer term, students involvement
- Partnership program (uo: 15% of indus. contrib.)
  - Governmental agencies (NSERC, Precarn, MITACS, Industry Canada) match an industrial contribution
- *Networks of Centres of Excellence / CFI*

There are no known *gold mine* IT licenses

J Strother Moore (University of Texas, Austin)

rali

SILC

SILC (*Système d'Identification de la Langue et du Codage*) automatically determines both the language in which a document is written, and the character encoding used. The current version recognizes close to thirty languages, and an average of three encodings per language.

It is now possible to integrate SILC in your application under the following systems: Windows, Linux, Solaris, MAC, HP-UX, AIX and SGI. SILC also exists in Java.

Liste of known languages and encodings

English cp1252
Chiness utf8
Japanese utf8
Spanish cp1252
German cp1252
Korean utf8
French cp1252
Italian cp1252
Portuguese cp1252
Dutch cp1252

Type some text in the box below, in the language of your choice. The system needs at least a few words to be able to produce a reliable identification

Ein Unternehmen zu gründen, geht nicht von heute auf morgen, sondern ist ein Prozess. Eine geniale Idee alleine reicht dabei nicht aus, um erfolgreich zu sein.

Submit text Choisir le fichier aucun sélectionné

Analyse Show details

rali

Czech cp1250
Czech iso-8859-2
Czech utf8
Danish cp1252
Danish cp850
Danish macintosh
Danish utf8
German cp850
German macintosh
German utf8

Ein Unternehmen zu gründen, geht nicht von heute auf morgen, sondern ist ein Prozess. Eine geniale Idee alleine reicht dabei nicht aus, um erfolgreich zu sein.

Submit text Choisir le fichier aucun sélectionné

The language is German, the encoding is utf8

Analyse Hide details

Ranking of candidates

Language	Encoding	Score
German	utf8	3.8506808
German	cp1252	4.559468
German	macintosh	4.559468
German	cp850	4.559468
Dutch	macintosh	9.825832
Dutch	cp850	9.825832
Dutch	cp1252	9.825832
Dutch	utf8	9.825832
English	cp1252	12.05258
English	utf8	12.05258

rali

Czech cp1250
Czech iso-8859-2
Czech utf8
Danish cp1252
Danish cp850
Danish macintosh
Danish utf8
German cp850
German macintosh
German utf8

Ein Unternehmen zu gründen, geht nicht von heute auf morgen, sondern ist ein Prozess. Eine geniale Idee alleine reicht dabei nicht aus, um erfolgreich zu sein.

Creating a company, does not happen overnight, but is a process. A brilliant idea alone is not enough to be successful.

Submit text Choisir le fichier aucun sélectionné

The language is German, the encoding is utf8

Analyse Hide details

Ranking of candidates

Language	Encoding	Score
German	utf8	3.8506808
German	cp1252	4.559468
German	macintosh	4.559468
German	cp850	4.559468
Dutch	macintosh	9.825832
Dutch	cp850	9.825832
Dutch	cp1252	9.825832
Dutch	utf8	9.825832
English	cp1252	12.05258
English	utf8	12.05258

rali

SILC : commercialization

- *?Que¿* - ALIS Technologies (1998 - 2001)
  - Binary and API in C (Java afterwards)
  - Integration into their web page translation service
  - Sold the whole system (source code and training corpora) to Oracle
- SILC - *Player*
  - COMScore, Mobikom, Excite
- Developed models for other languages

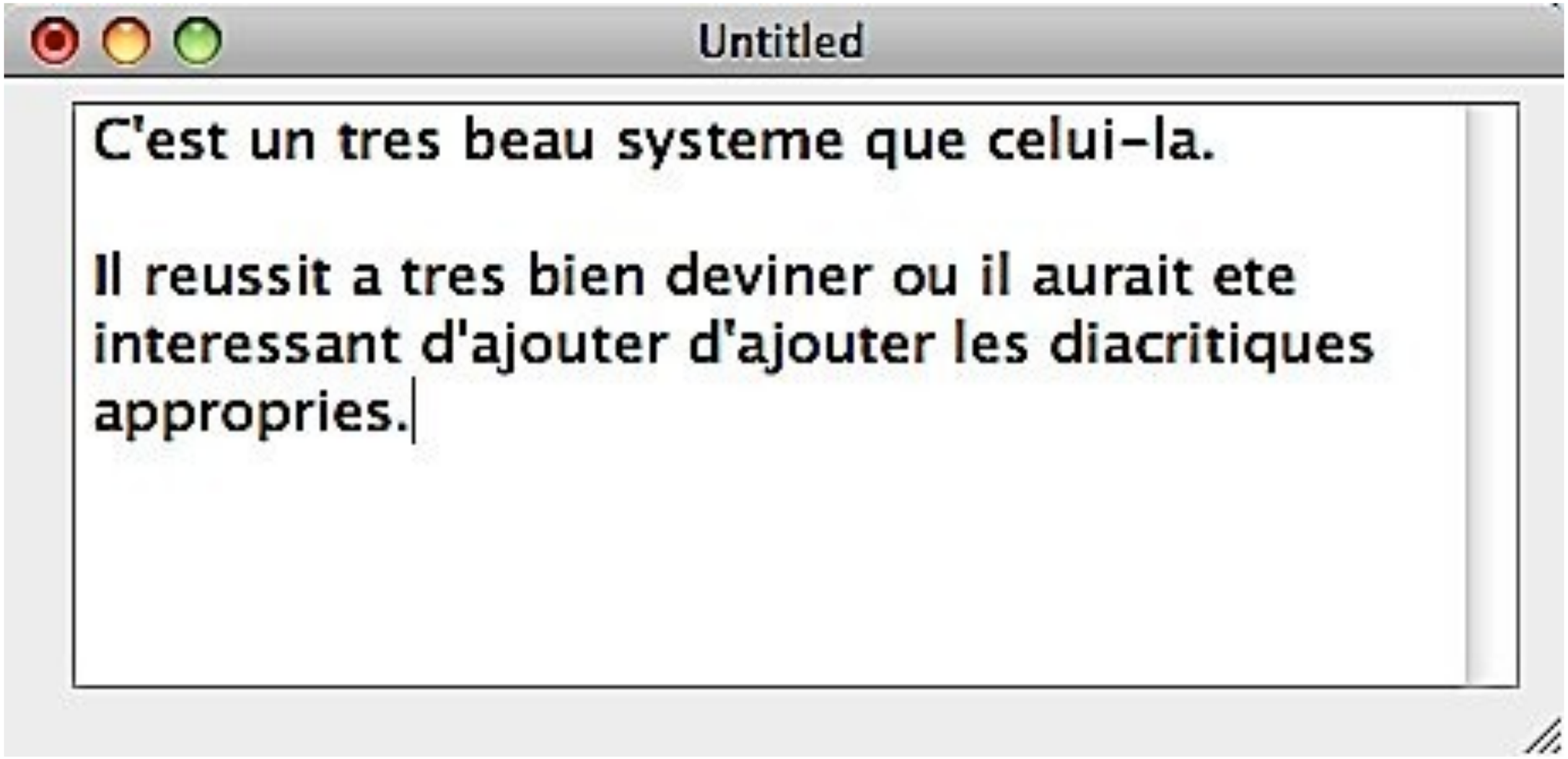
rali

SILC : lessons learned

- Simple idea (in principle!)
- SILC apply vs SILC train-merge
- Licensing rights were long into making
  - no single year agreements, please...
  - payment schedule not always respected
- Simpler with a good licensing model
- Language recognition is now *given* in most text editors and MT web pages

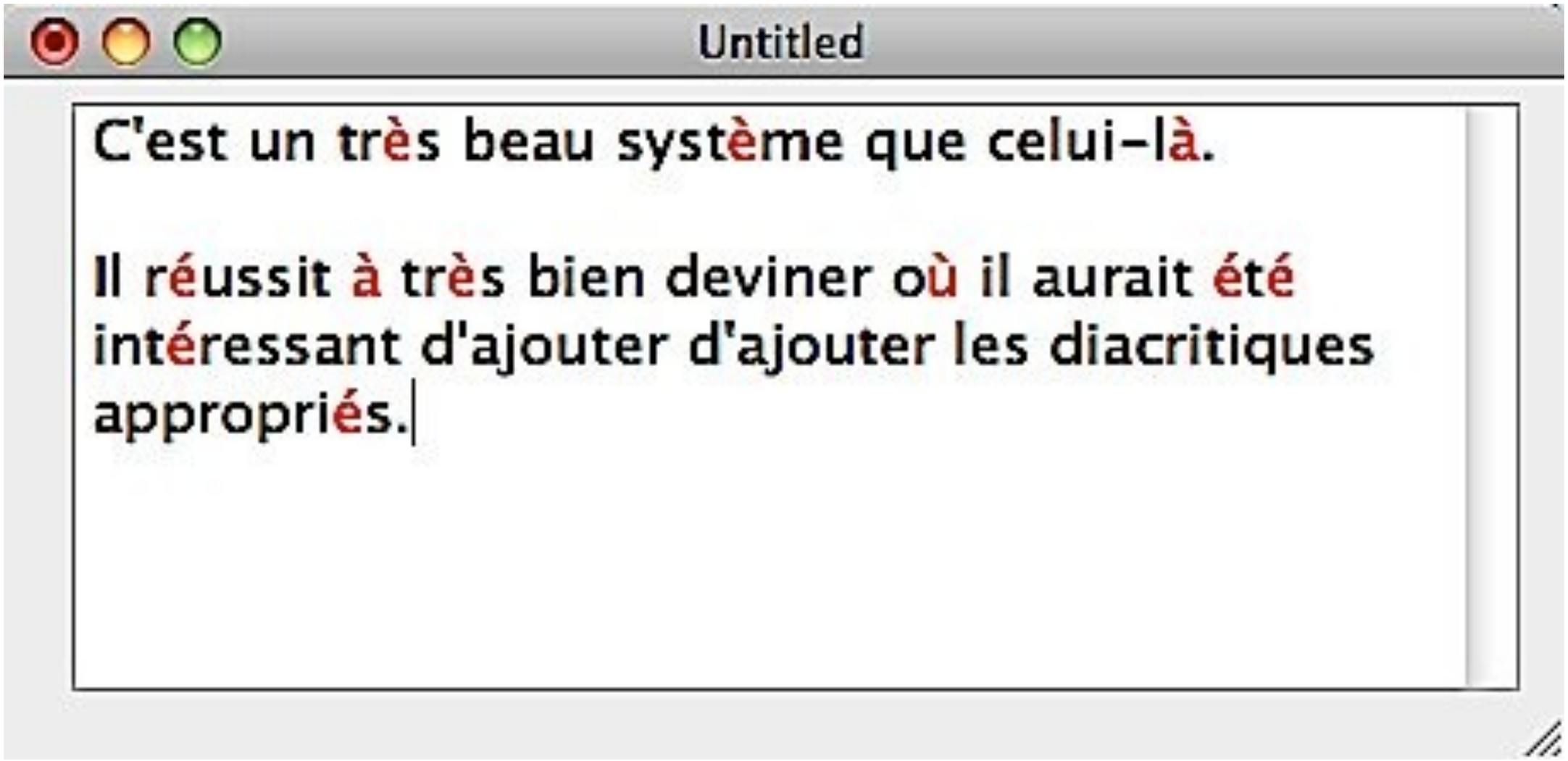
rali

Réacc: French Accentuation



rali

Réacc: French Accentuation



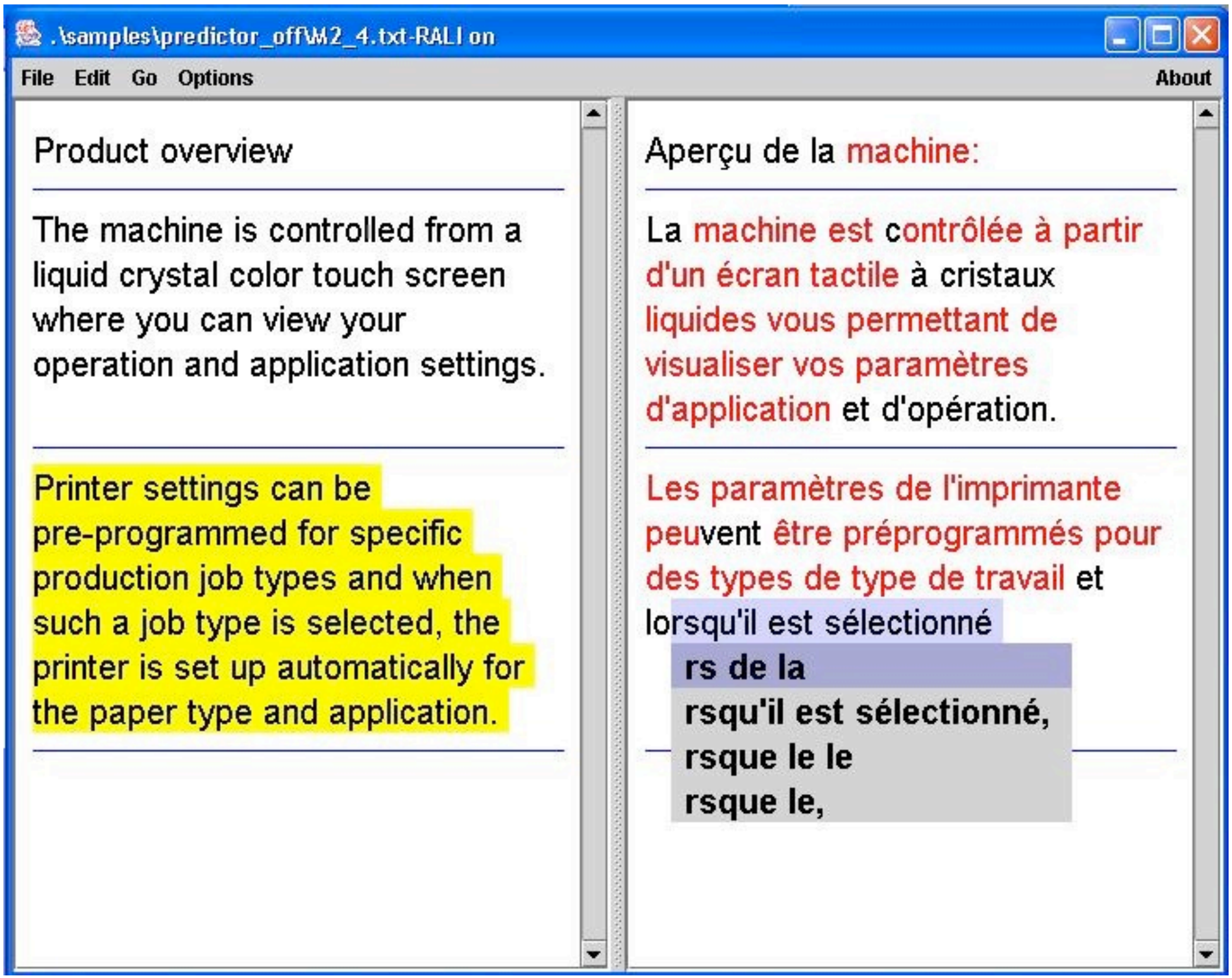
rali

Réacc

- Never managed to sell any license !!!
- Nice for demos, student projects and scientific articles but no market
  - Word plug-in
  - Mac application
- Also for Vietnamese

rali

TransType



rali

Probabilistic model

$$p(t|t',s) = \underbrace{p(t|t')\ \lambda(\Theta(t',s))}_{\text{langage}} + M\ \underbrace{p(t|s)\ [1-\lambda(\Theta(t',s))]}_{\text{traduction}}$$

source sentence: **They are getting into seeding.**
s

ongoing translation: **Ils vont c** *ommencer*
t'

prefix
completion

active vocabulary V

actuellement
agriculteurs
amorcer
commencent
commencer
cultivateurs
elles
ensemencement
ils
leurs
semailles
semences
sont

p(commencent | they are getting into seeding, 3) = 0.3
p(commencer | they are getting into seeding, 3) = 0.2
p(cultivateurs | they are getting into seeding, 3) = 0.1

translation probabilities

λ

p(commencent | ils vont) = 0.001
p(commencer | ils vont) = 0.1
p(cultivateurs | ils vont) = 0.0005

language probabilities

rali

TransType milestones

- TransTalk (CITI, 1995)
- Target-Text Mediated Interactive Machine Translation (Foster, 1997)
- Strategic grant (1997-2001) industrial partner: Machina Sapiens
- TransType2 (2002-2005) 5th Framework European Project IST-2001-32091

- F. Casacuberta, J. Civera, E. Cubel, A. L. Lagarda, G. Lapalme, E. Macklovitch and E. Vidal.
  Human interaction for high-quality machine translation.
  Communications of the ACM, vol. 52, number. 10, p. 135-138, oct 2009

rali

TransType2 Partners

- Canadian (CRSNG-MDER)
  - RALI - Université de Montréal
  - Société Gamma (Translation firm - *sister firm* of Terminotix)
  - CRTL
- European (CEE)
  - Atos Origin (Spain)
  - University of Aachen (Germany)
  - University of Valencia (Spain)
  - Xerox Research Centre Europe (France)
  - Celer Soluciones (Translation firm)

rali

Experimentations in TransType2

- 3 language pairs
  - English↔{French, Spanish, German}
- 3 translation model (RALI,RWTH,ITI)
- 2 types of texts
  - technical
  - parlementary (canadian et EC)
- Vocal interface
- Confidence estimation
- Adaptation

rali

TransType2 - Evaluation

- reduction in the number of keys typed
- *substitute* production measure
- *theoretical evaluation* predicts 60% of saving
- *pratical evaluation* on the order of 30%
- translators liked the idea
  - suggestion less intrusive than translation memory
  - no spelling error in suggestions
  - but it induces a literal translation

rali

Lessons learned

- cognitive load for the translators
  - a single long suggestion is better than many shorter ones
- should limit the number of choice and frequency of appearance
- acceptance = typing 3 chars
- but no industrial application
  - ahead of its time
  - ATOS did not do its job

rali

TransSearch

Bookmark TransSearch
(what is this?)

Document collection : House of Commons Hansard (1986-2010)

Expression: in keeping with Search Bilingual query

1	On leur donne des responsabilités moindres ? mieux adaptées à leurs capacités, j'imagine ?, mais ils continuent de faire partie du Cabinet, avec tous les avantages qui viennent avec.	They get moved to lesser responsibilities, more **in keeping with** their capabilities, I suppose, but they stay in cabinet with all of the perks that go with being in cabinet.
2	Il examinera ces renseignements pour ensuite faire rapport à la Chambre ou au comité compétent, ce qui est tout à fait conforme aux pratiques courantes en ce qui a trait à ce genre de documents.	That information will be reviewed by him. There will be a report back to the House or to a proper committee. This is very much **in keeping with** practices around documents such as these.
3	Cela détonne dans un pays qui veut, et dit vouloir, investir dans ses habitants.	That is not **in keeping with** a country that is looking forward and saying that it wants to invest in its people.
4	Conformément à ces priorités, le budget mène à terme notre Plan d'action économique.	**In keeping with** these priorities the budget completes our economic action plan.
5	Dans le cadre de cette révision, on en profitera sûrement pour éliminer des programmes qui gênent le gouvernement, qui ne sont pas conformes à son idéologie.	When doing their review, they will certainly take advantage of it to eliminate programs that annoy the government, that are not **in keeping with** its ideology.

TransSearch - 1

- Initially (1992-1995) on Sun workstations
- Web version 1997
  - Hansard / Court Rulings 1986-1994
  - Free (as in *free beer*) for everybody
- Tried hard to convince industry to develop a *Web service* with this idea

rali

TransSearch - 2

- RALI launched its own commercial service in 2001
  - updated corpora (Hansard and Court Rulings)
  - revamped user and administration interface in Perl
  - kept the same textual data base (MG)
  - developped a business model (SAAS before its time)
- In two years
  - 1 500 licenses (1000 at the Translation Bureau)
  - Enough to pay the development and maintenance

rali

TransSearch - 3

- Deal with Terminotix in 2003
  - administration and daily update of Hansard
  - recruiting new customers
  - service agreement
    - for adding new corpora
    - maintaining the server software
  - added Spanish-English, Spanish-French corpora
- Three-year deal renewed since
- Win-win situation

rali

TransSearch - 4

- After years of *evolution*
  - hard to maintain and non incremental indexing
  - new research : translation spotting and its applications
- NSERC Collaborative Research and Development Grant
  - develop translation spotting
  - study new applications using *transspots*
  - embed these into a new system

rali

Expression: in keeping with

Collections: Hansard 1994-2006

Search | Bilingual Query

165 translations of *in keeping with* in 670 occurrences

conforme à	156
conformément à	106
respecte	48
correspondant à	29
en conformité avec	16
dans le sens de	14
fidèle à	13
dans le respect	12
à fait	10
compte tenu	9
compatible avec	9
en fonction	9
répondait	8
inscrit dans	7
dans le cadre de	6
en vertu de	6
dans l'esprit de	6
en logique avec	4
étant donné	4
qui va dans le sens d'	4
contraire	4
partie des	4

conforme à 156

Fourth, the judge must be convinced that a conditional sentence is **in keeping with** the general principles of proportionality of the sentence.	Quatrièmement, le juge doit être convaincu que l'emprisonnement avec sursis est **conforme** aux principes généraux de la proportionnalité de la peine.
It is a bill that is **in keeping with** the campaign and election commitments made by the Conservative Party of Canada to Canadians.	Ce dernier est **conforme aux** engagements pris par le Parti conservateur pendant la campagne électorale.
All this is **in keeping with** our Canadian values of compassion and generosity in times of need, a quality displayed across every community in Canada.	Voilà qui est **conforme à** nos valeurs canadiennes de compassion et de générosité qui s'expriment dans des cas comme celui-là, une qualité commune à toutes les collectivités du Canada.
The reserve was **in keeping with** the established budgetary practice of setting aside policy reserves for specific contingent purposes.	La réserve était **conforme à** la pratique budgétaire établie qui consistait à mettre de l'argent de côté pour certaines éventualités.
I am certain that passing this bill, **in keeping with** Canada's policy of equal treatment of the parties, will contribute to reopening the roadmap to peace and will ensure a lasting peace between Israel and Palestine.	Je suis convaincu que l'adoption de ce projet de loi, **conforme à** la politique canadienne de traitement égal des parties, contribuera à relancer la feuille de route pour la paix et fera en sorte que la région connaîtra une paix durable au bénéfice d'Israël et de la Palestine.
When the government introduced legislation, specifically Bill C-31 and Bill C-32, since, as explained the hon. Parliamentary Secretary to the Government House Leader, it was as a complement **in keeping with** ... Canadian practice...to confirm major changes in government organization through legislation .	Lorsque le gouvernement a déposé les projets de loi C-31 et C-32, comme l a expliqué le secrétaire parlementaire du leader du gouvernement à la Chambre, c était à titre de mesure complémentaire, **conforme à** la pratique canadienne, pour « confirmer tout changement d importance dans l organisation gouvernementale ».
I would simply ask that we preserve the current definition of marriage since it is wholesome for the common good, **in keeping with** the natural law and in conformity with God's design for the world.	Je demande simplement que nous préservions la définition actuelle du mariage, puisque le mariage est salutaire pour le bien commun et qu'il est **conforme** à la loi naturelle et au dessein de Dieu pour les êtres humains.
This greater concentration in Africa is **in keeping with** Canada's	Mettre ainsi l'accent sur l'Afrique est **conforme à** l'engagement

Automatic judgement summarization

- Area in which clients pay for summaries
- Usual approaches do not work well
- Extraction is preferable to abstraction
- A. Farzindar thesis (2005 - U de Montréal)
  - thematic segmentation
  - tabular presentation

rali

NLP Technologies Inc. launched by A. Farzindar

- Web Services
  - Decision Express
  - Biblio Express
  - Statistic Express
  - Search Express
- *High volume* domains: immigration, tax, intellectual property
- Used by many lawyer offices and by judges from the Federal Court

rali

ASLI

Automatic Summarization of Legal Information

June 2007-2008

- Industrial collaboration (Precarn)
  - NLP Technologies
  - RALI
  - Lawyers
- Automatic Summarization (French and English)
  - Finite state automata
  - Vocabulary separation from identification rules
- Machine Translation experimentation
- June 2007-2008

rali

ISASLI

Intelligent system for **S**emantic processing, **A**utomatic translation and **S**ummarization of **L**egal **I**nformation

(Jan-Dec 2009)

- Revision help - RevSum
- Statistical summarization
- Machine translation (another type of court CSST)

rali

ASLI-ISASLI
Fruitful collaboration

- motivated industrial (entrepreneurial student !)
- willing to try new ideas
- willing to publish results
- MT proved to be a very successful *side effect*

rali

Druide informatique
Statistical filtering of spelling corrections

- French spelling checker - Antidote
  - Excellent symbolic parser of French
  - tradeoff between noise and silence
  - clients are sometimes annoyed by too many suggestions
- 4 months contract
- Decision trees for 14 types of spelling errors:

  `là` vs `la`, `er` vs `ez`, verb modes, etc.
- Antidote HD (sept 2009) implements 8 classifiers

rali

Other industrial partners

- Bell University Laboratories (1999-2003)
  - Automatic Investors E-mail processing
  - NSERC-RDC
    - 2 PhD + 2 MSc
- SOQUIJ & Lexum
  - anonymization of judgements (NOME)
- Environment Canada (2009-2011)
  - MITACS seed project for Multiformat Environmental Information Dissemination

rali

What happened ?

- SILC vs Reacc
- TransType
- TransSearch
- ASLI - ISASLI
- Druide

What happened ?

- SILC vs Reacc
- TransType
- TransSearch
- ASLI - ISASLI
- Druide

Good research ideas are not enough !

rali

Faculty, be aware that

Faculty, be aware that

- Need an internal champion within the industry

Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company

Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects

Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a *hurry*

rali

Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a *hurry*
- IP discussions between *lawyers* can be long and arduous

rali

Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a *hurry*
- IP discussions between *lawyers* can be long and arduous
- Big companies are not easier to work with

rali

Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a *hurry*
- IP discussions between *lawyers* can be long and arduous
- Big companies are not easier to work with
- Not a *pot of gold*

rali

Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a *hurry*
- IP discussions between *lawyers* can be long and arduous
- Big companies are not easier to work with
- Not a *pot of gold*

Wait until you are tenured !!!

rali

But

- Fun and stimulating
- Reality check
- Once industry decides to put *real* money, industrial grants are relatively easier to get than pure research grants

rali

What I wish industry would *understand*

- scientific publications and conferences are good publicity
- the best benefits of a project is a privileged access to the students that worked on the project
- Each dollar spent in outside research should be matched by at least a dollar within the company

rali

References

- J Strother Moore, Lawrence Snyder, Philip A. Bernstein, University-Industry Sponsored Research Agreements, CRA - Best Practice Memo, 2003.
- H. Charmasson, J. Buchaca, N. Milton, D. Byron, Canadian Intellectual Property Law for Dummies, Miltons IP, 2009.

rali

Questions ?

Comments ?

rali