SNLP 2007, Pattaya, 15/12/07

Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a "linguistic survival" system under construction

Christian Boitet[1], Pushpak Bhattacharyya[2], Etienne Blanc[1], Sanjay Meena[2], Sangharsh Boudhh[2], Georges Fafiotte[1], Achille Falaise[1], Vishal Vacchani[2]

[2] Centre For Indian Language Technology (CFILT)
Indian Institute of Technology (IIT Bombay), Mumbai- 400 076, India
e-mail : {pb,sanjay,sangharsh,vishalv}@cse.iitb.ac.in

[1] GETALP (Study Group for Translation and Processing of Languages and Speech)
LIG (Grenoble Informatics Laboratory), UJF (Univ. Grenoble 1) & UPMF (Univ. Grenoble 2)
e-mail : {Name.Surname}@imag.fr

Outline

- Introduction
- 1 Objectives of building SurviTra
- 2 Corpus building
- 3 Dictionary construction, with UWs
- Conclusion & perspectives

Context

- Cooperation with India (IIT-Bombay, Pondichéry)
  - VTHFraDial 2003-05
    - ERIM-collect of spoken bilingual dialogues
    - French↔Vietnamese, Tamil, Hindi (Chinese done before)
  - CIFLI (ARCUS programme) 2006-08
    - Further dialogue collection
    - Data collection for crosslingual communication
      - ✓ 3-4 Indian languages targeted: Hindi, Marathi, Tamil, Konkani
    - Prototype MT by UNL between French & these languages
- Concretely
  - 2x2-m + 2x1-m stays by Profs (Bhattacharyya, Pannirselvame)
  - 2x2-m + 2x3-m stays by Master students (Sangarsh, Sanjay)
  - 5+3-w + 2x3-m work in Grenoble (Blanc, Fafiotte, Boitet, Falaise)

Objectives of building SurviTra

- Practical and "political" goals
  - Survival Translation abroad
  - Use a PC in an Internet café in "deep India"
  - Showcase for MT
    - But be safe: 10% MT, 90% phrasebook!
    - However: 60% = 2/3 of 90% with "variables and variants"
      - Please give me $nb $fruit juice
      - $nb [verre(s)] de jus de $fruit → 1 verre de jus d'ananas
- Functional goals
  - Make it an "extended multilingual chat"
  - Contributions from users to phrasebook, dictionary
- Research goals
  - Use "UNL graphs with variables" to generate variants
  - Unify "modernized UW sets" (IITB + UPM/IPPI 300,000 U++C/WN)

Specification of extended ML tchat

Survivor French	Helper Indian Language	Other English	UNL graphs
t1: S1	t1:	t1:	t1:
t2:	t2: S1	t2: S1	t2: U1
t3:	t3: S2	t3:	t3:
t4: S2	t3:	t4: S2	t4: U2
Correction in French		Chat pane	
t5: S2'	t5:	t5:	t5:
t6:	t6: S2'	t6: S2'	t6: U2'
t7: S3	t7:	t7:	t7:
	t8: S3	t8: S3	t8: U3
…	…	…	…

Phrasebook					Dictionary			
French	H/M/T	E	UNL (opt)		French	H/M/T	E	UNL
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW
…			Search & contribution pane					
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW

Navigation in phrasebook	Display controls	Modes	Timing	Statistics
		Control pane		MT usage
Input area				

From Koine to SurviTra

- Koine: ML tchat by A. Falaise (Ph.D + Prosodie Inc.)
  - Multimedia targeted Speech, text
    - But current state-of-the-art ASR unusable on the phone
  - Then multilingual tchat with helps ot bilingual tchat
    - Lexical support, MaT with interactive disambiguation
  - and some innovations
    - Past turns editable many turns are for correction!
    - Data collection integrated 20Mb collected in 2004
- SurviTra: Futher extensions
  - 2 languages
  - + English even if only 15% Indians are "operational" in English
    - (less French?)
  - + UNL small icons, click to display graph
    - Used only by "expert" contributors
    - But usage and contribution environments should be integrated

Screen shot of Koine-SurviTra

- To be added

Architecture of SurviTra

- Resource components
  - multilingual translation memory
  - multilingual lexical database
  - domain hierarchy
  - repertory of users and profiles
  - database containing information (metadata) on usable MT systems & various statistics.
- Software components
  - 3 web services adapted to support first-level functions
    - ✓ search, actual usage ; addition or correction of fixed sentences
    - multilingual chat facility (Koine by A. Falaise);
    - web contributive translation support (BEYTrans by Y. Bey)
    - multilingual database for MT (PIVAX by H.T. Nguyen)
      - ✓ See yesterday's talk
  - Three more components be integrated, to
    - handle NL and UNL templates
    - call MT systems on "not found" sentences (<10%)
    - edit UNL graphs.

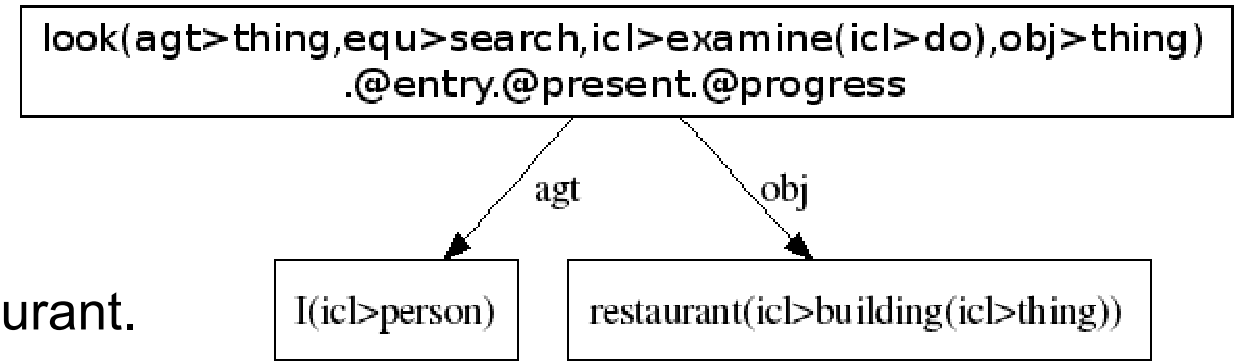
Corpus building

- Languages
  - French, Hindi, Marathi, English + UNL
- Domains / situations targeted now
  - speaking with authority figures
  - accommodations
  - money matters, problems (police, theft)
  - common questions
  - restaurants
  - shopping
  - travel
  - sentences for acceptance, refusal, or salutations
- Extensions (after CIFLI)
  - languages and domains
  - depth within each domain

Example

English: I am looking for a restaurant.

UNL:



[S:23]

;I am looking for a restaurant.

{unl}

agt(look(icl>examine>do, equ>search, agt>thing, obj>thing).@present.@progress.@entry,
I(icl>person))

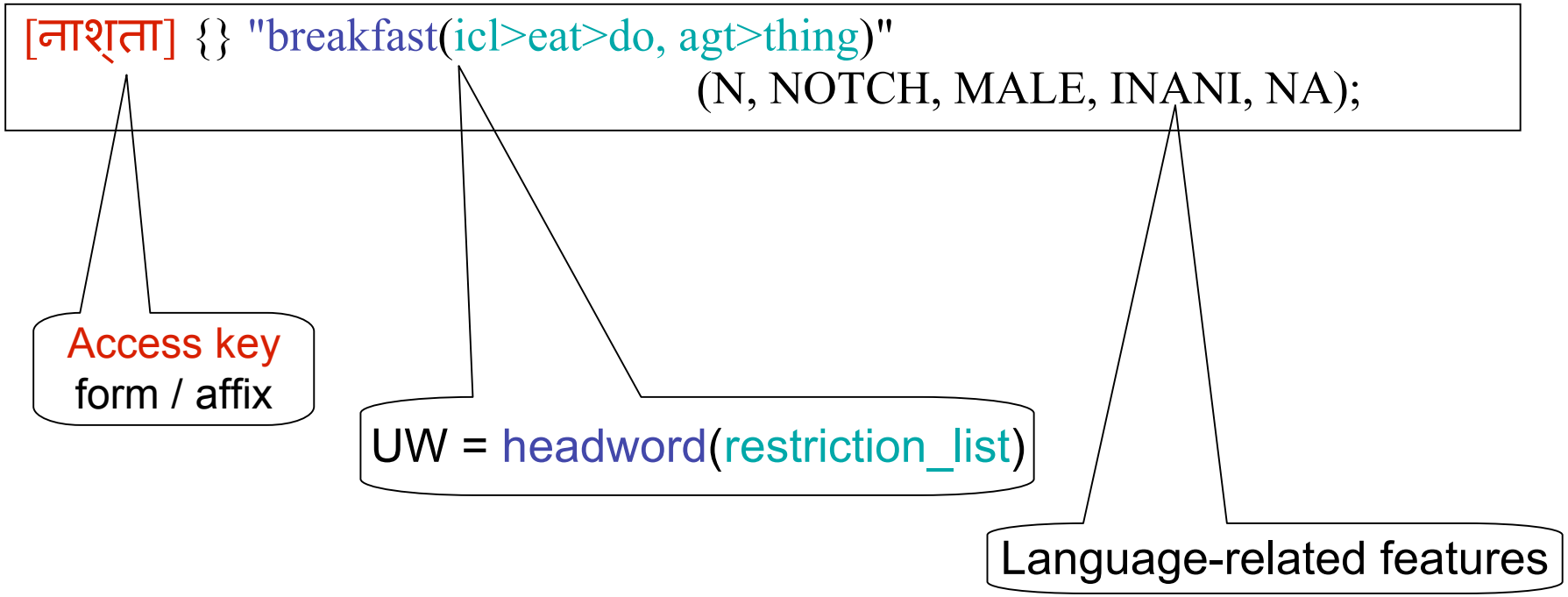
obj(look(icl>examine>do, equ>search, agt>thing,obj>thing).@present.@progress.@entry,
restaurant(icl>building>thing))

{/unl}[/S]

Hindi: मैं एक रेस्टोरेंट खोज रहा हूँ।

French: Je cherche un restaurant.

Hindi-UNL dictionary entry



Dictionary construction, with UWs

- Specifications
- Methodology
- Unification of UW dictionaries
  with U++C standards

Methodology

- Existing sources used
  - Hindi phrasebook from wiki-travel — http://wikitravel.org/
  - Hindi-French *Green Book* — for restaurant domain
  - UNL-Hindi Deconversion Dictionary — CFILT (IITB)
    > 116,000 entries
  - U++C UW website
    ≈ 300,000 UWs built by
    - ✓ using previous UW sets
    - ✓ Enriching UWs by constraints based on WordNet:
      icl>WN_synonym
- Missing Hindi lexemes and UWs taken from:
  - Hindi-French lexicon — paper
  - UNL-French lexicon — GETAlp

A Hindi-French lexicon entry

UW	English	Hindi	French
toast(icl>bread>thing)	toast	टोस्ट	des toasts

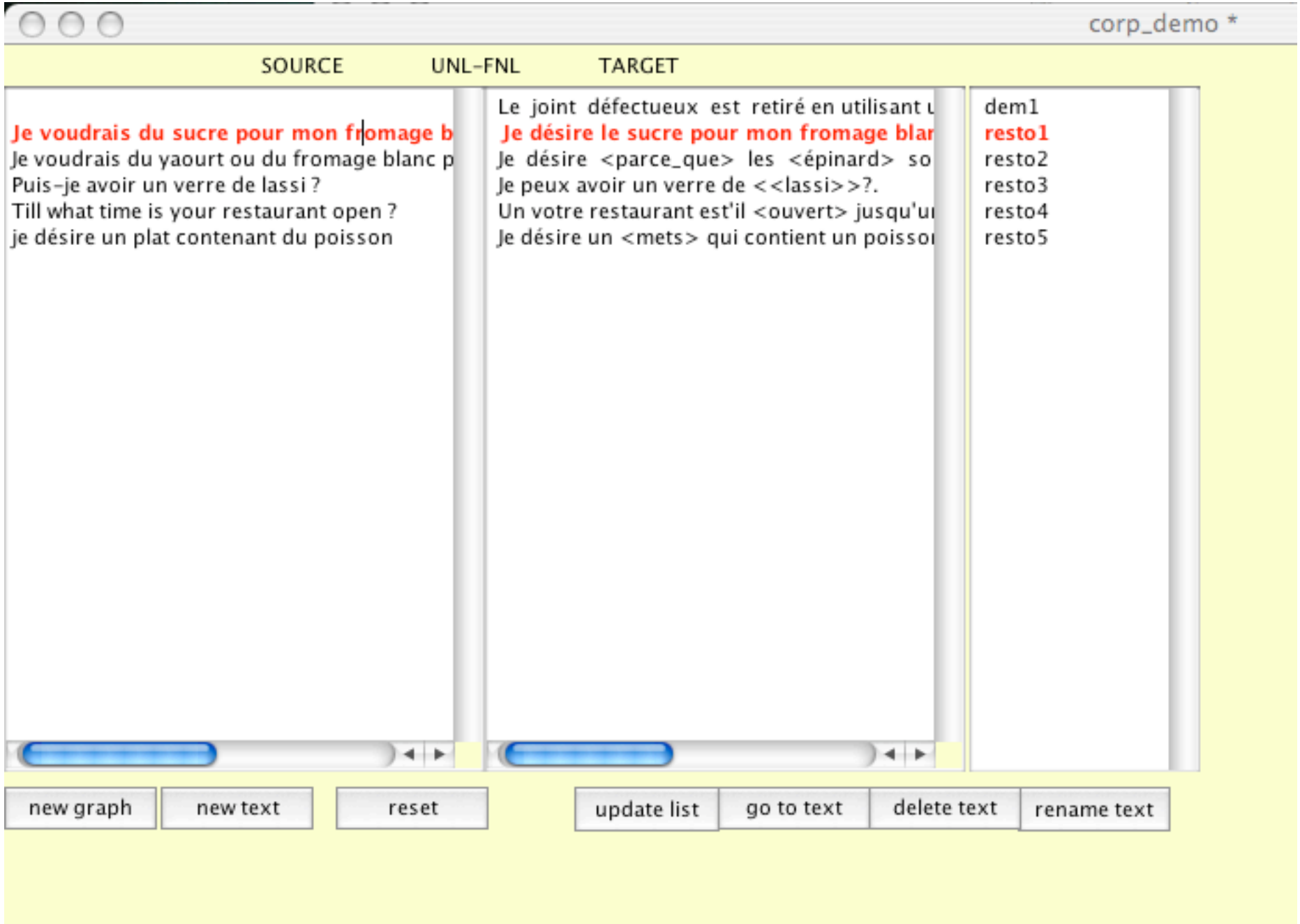
Phrasebook construction

Use of Revolution "stacks"

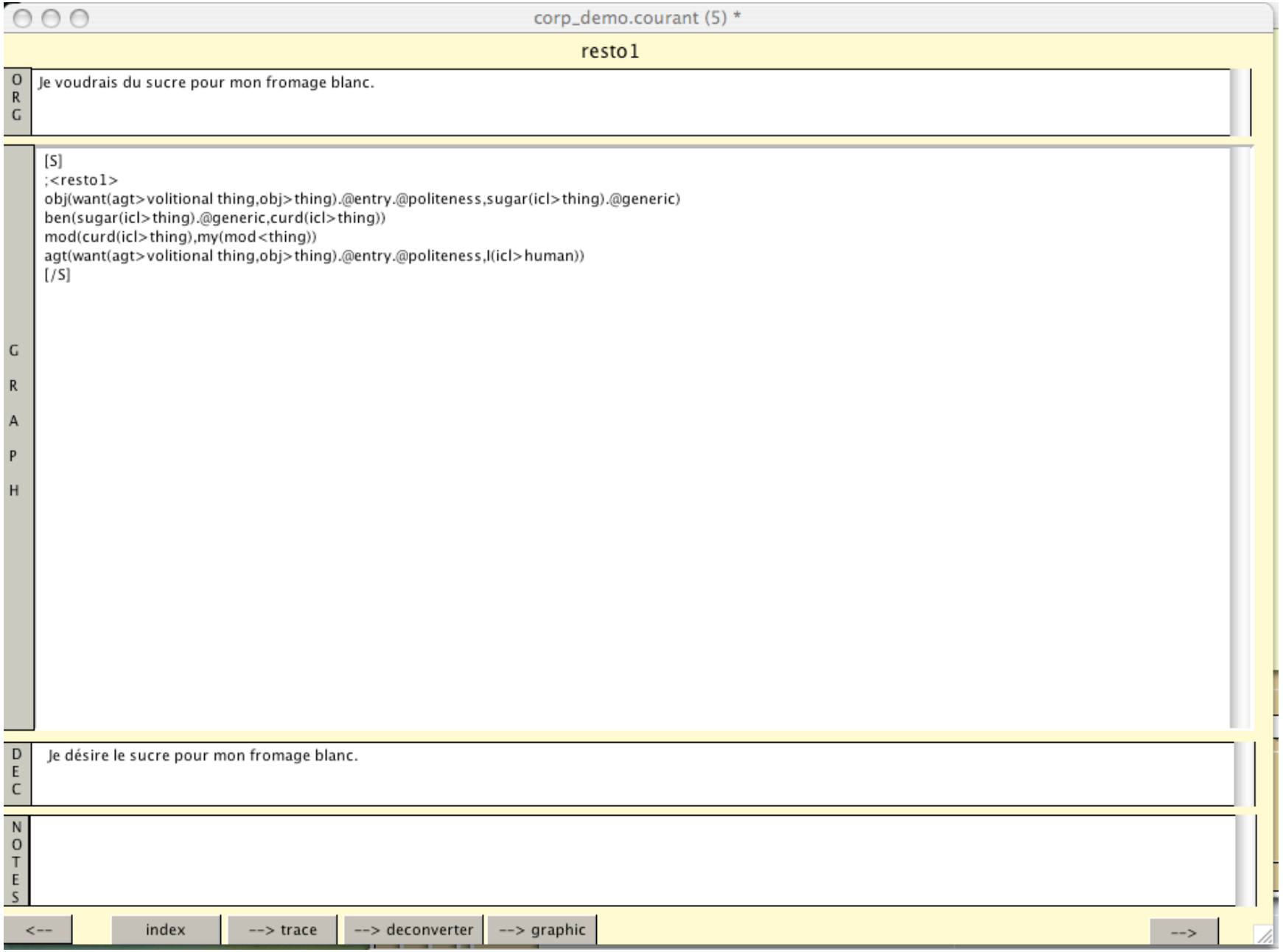
- Compatible HyperCard, portable — E. Blanc

UNL graph images produced by unldeco server

- Parser + displayer → image files — G. Sérasset


corp_demo *
SOURCE
UNL-FNL
TARGET
Je voudrais du sucre pour mon fromage b
Je voudrais du yaourt ou du fromage blanc p
Puis-je avoir un verre de lassi ?
Till what time is your restaurant open ?
je désire un plat contenant du poisson
Le joint défectueux est retiré en utilisant u
Je désire le sucre pour mon fromage blar
Je désire <parce_que> les <épinard> so
Je peux avoir un verre de <<lassi>>?.
Un votre restaurant est'il <ouvert> jusqu'u
Je désire un <mets> qui contient un poisso
dem1
resto1
resto2
resto3
resto4
resto5
new graph
new text
reset
update list
go to text
delete text
rename text

I would like sugar for my cottage cheese

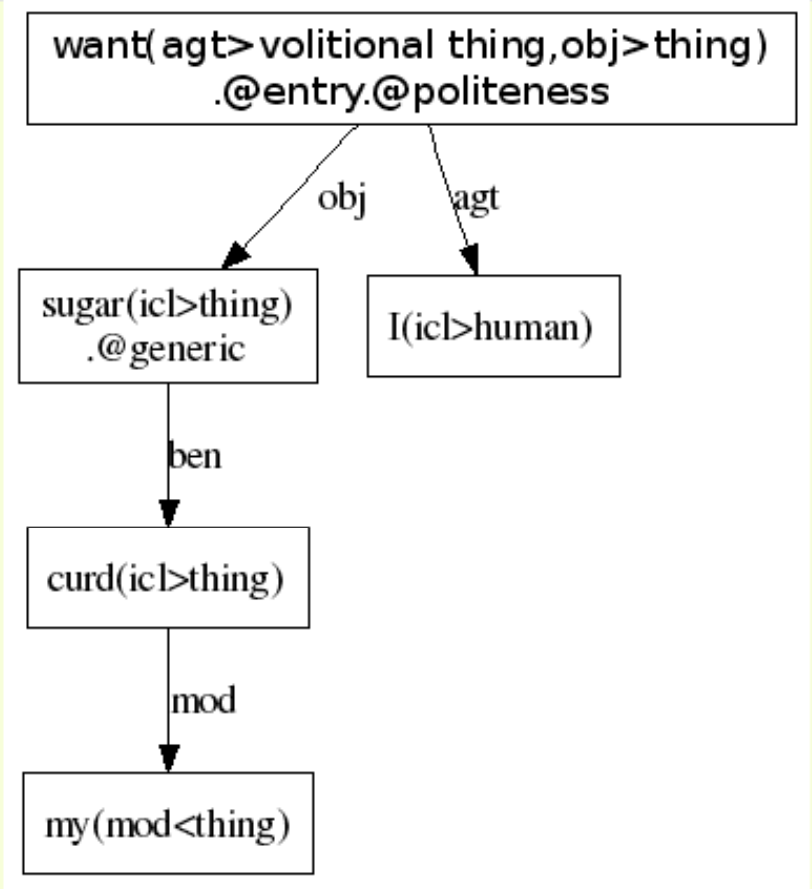


corp_demo.graphic (20) *

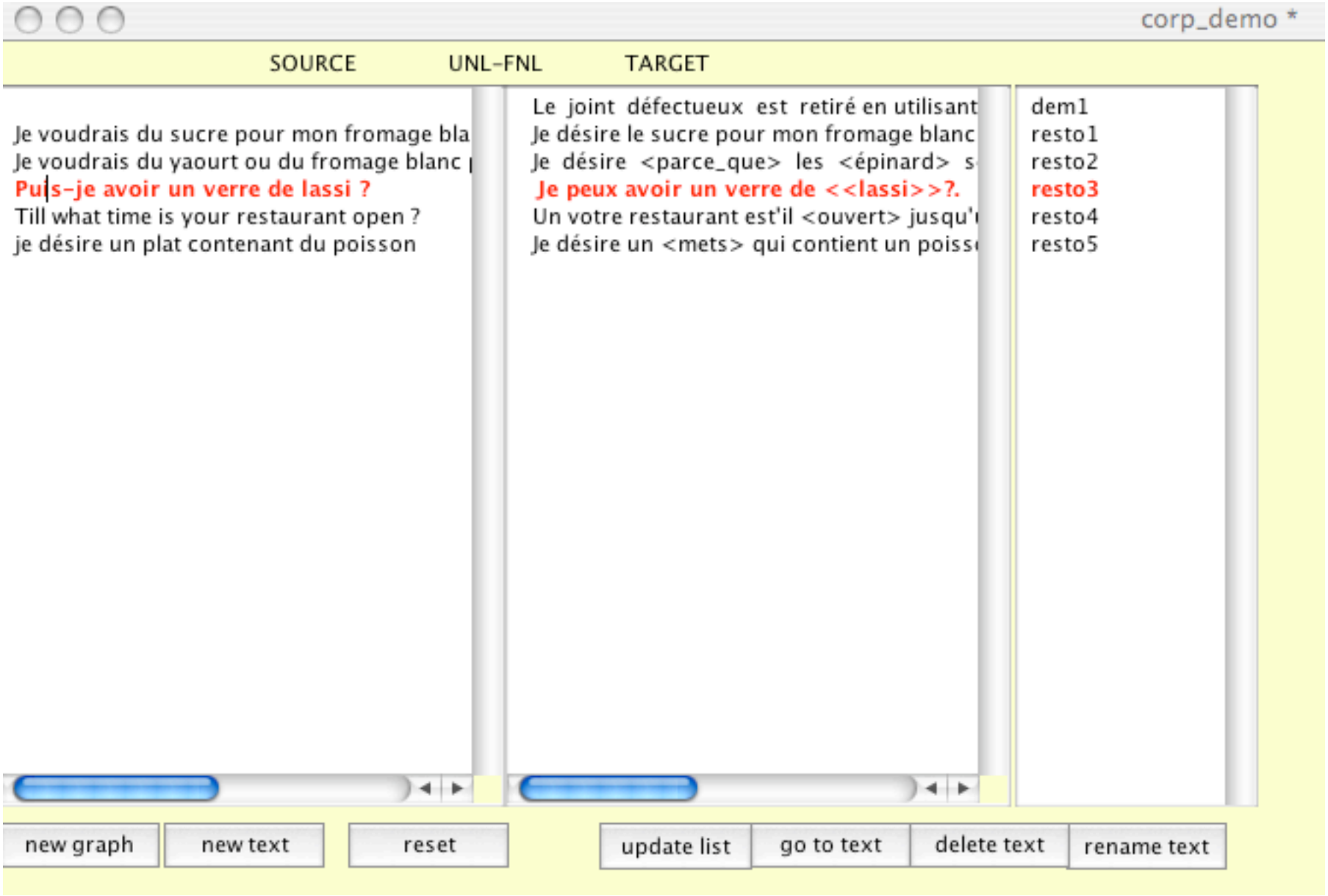
to

resto1

Je voudrais du sucre pour mon fromage blanc.



index | --> trace | --> deconverter | --> graphic | -->


corp_demo *
SOURCE
UNL-FNL
TARGET
Je voudrais du sucre pour mon fromage bla
Je voudrais du yaourt ou du fromage blanc
Puis-je avoir un verre de lassi ?
Till what time is your restaurant open ?
je désire un plat contenant du poisson
Le joint défectueux est retiré en utilisant
Je désire le sucre pour mon fromage blanc
Je désire <parce_que> les <épinard> s
Je peux avoir un verre de <<lassi>>?.
Un votre restaurant est'il <ouvert> jusqu'
Je désire un <mets> qui contient un poiss
dem1
resto1
resto2
resto3
resto4
resto5
new graph
new text
reset
update list
go to text
delete text
rename text

corp_demo.courant (3) *

resto3

ORG

Puis-je avoir un verre de lassi ?

GRAPH

[S]
;<resto3>
agt(may(equ>might).@entry.@interrogative,I(icl>human))
obj(may(equ>might).@entry.@interrogative,get(equ>obtain(icl>do,agt>thing,obj>thing)))
obj(may(equ>might).@entry.@interrogative,glass(icl>container))
mod(glass(icl>container),lassi(icl>food))
[/S]

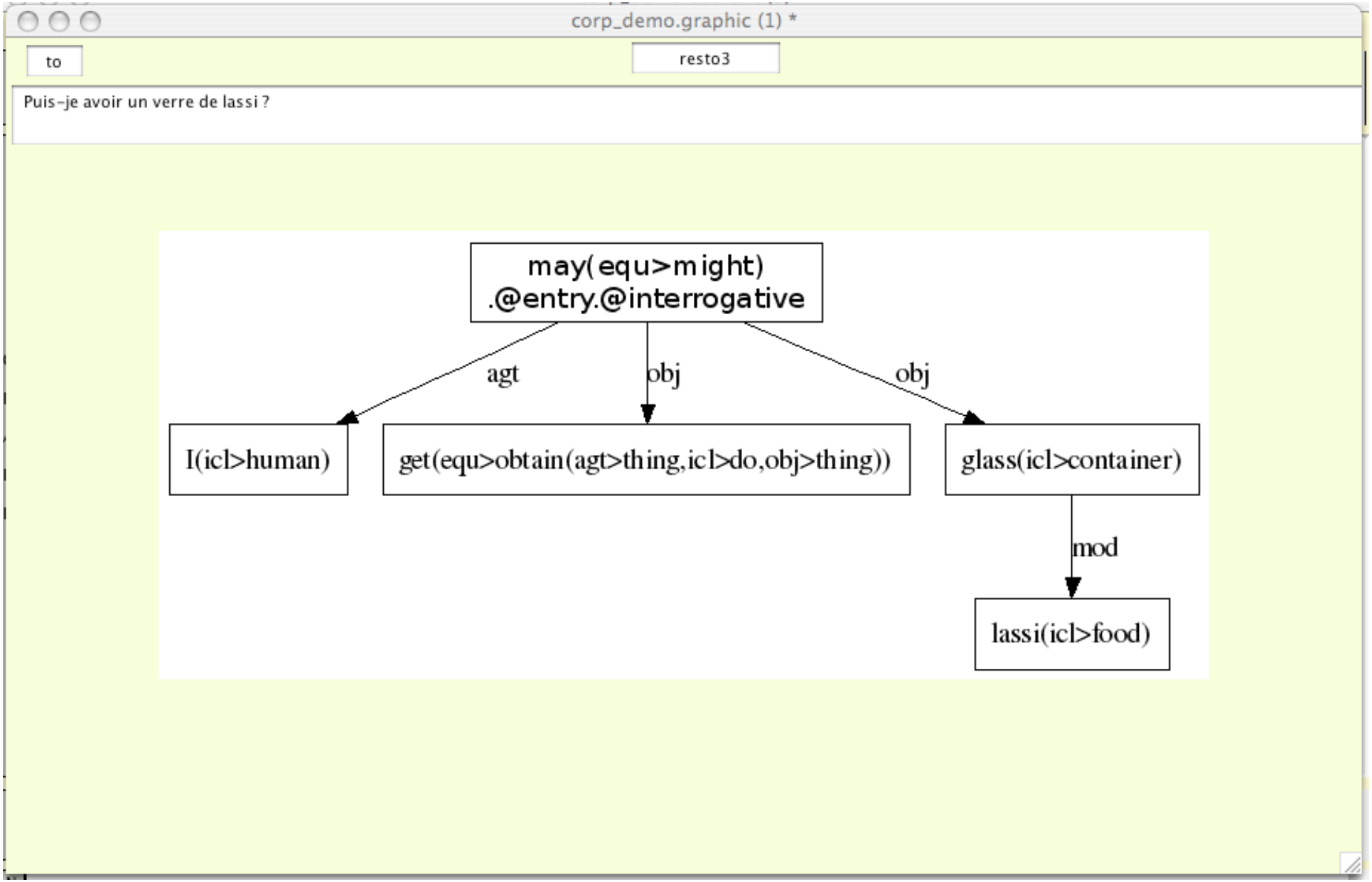
DEC

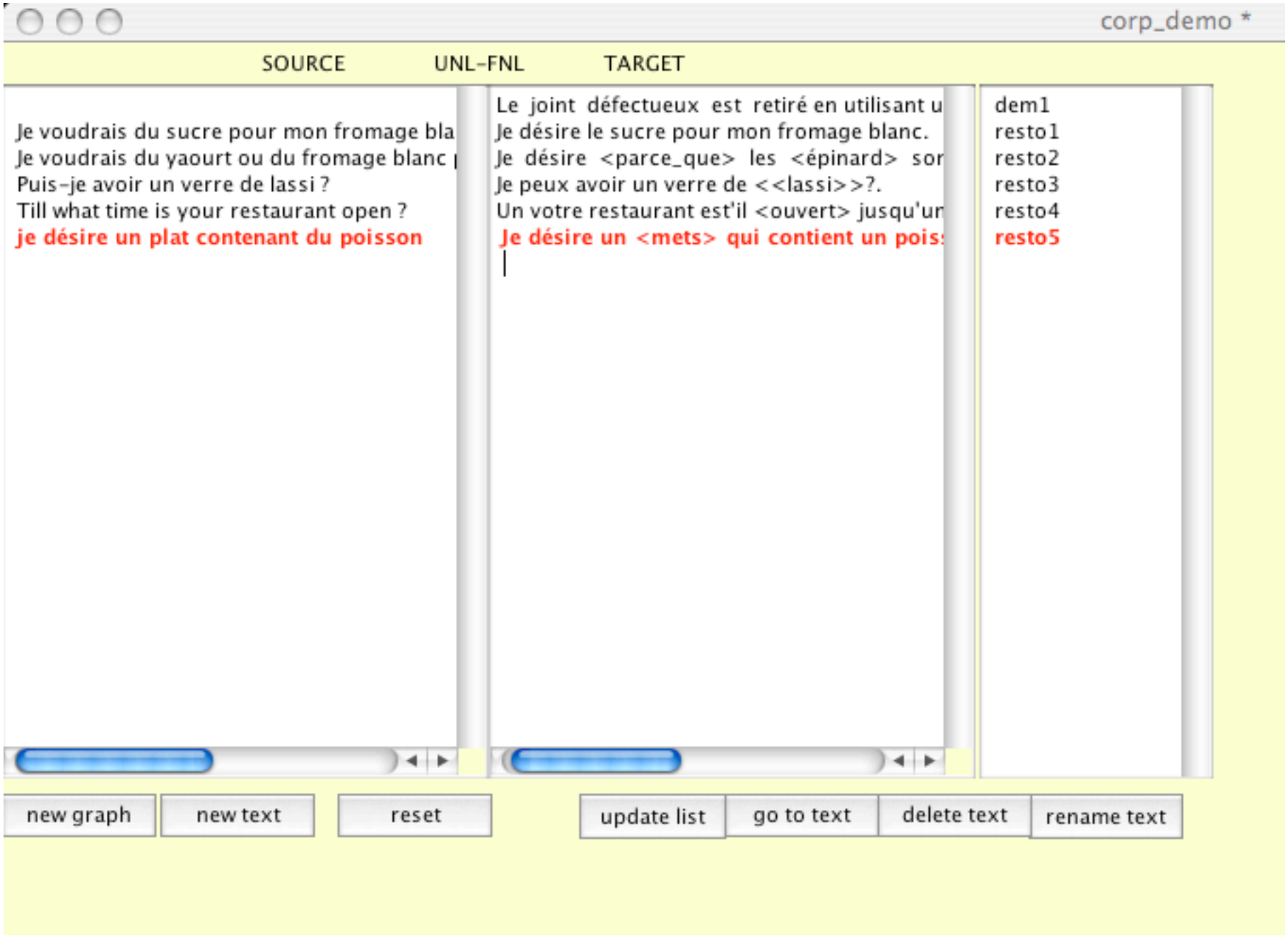
Je peux avoir un verre de <<lassi>>?.

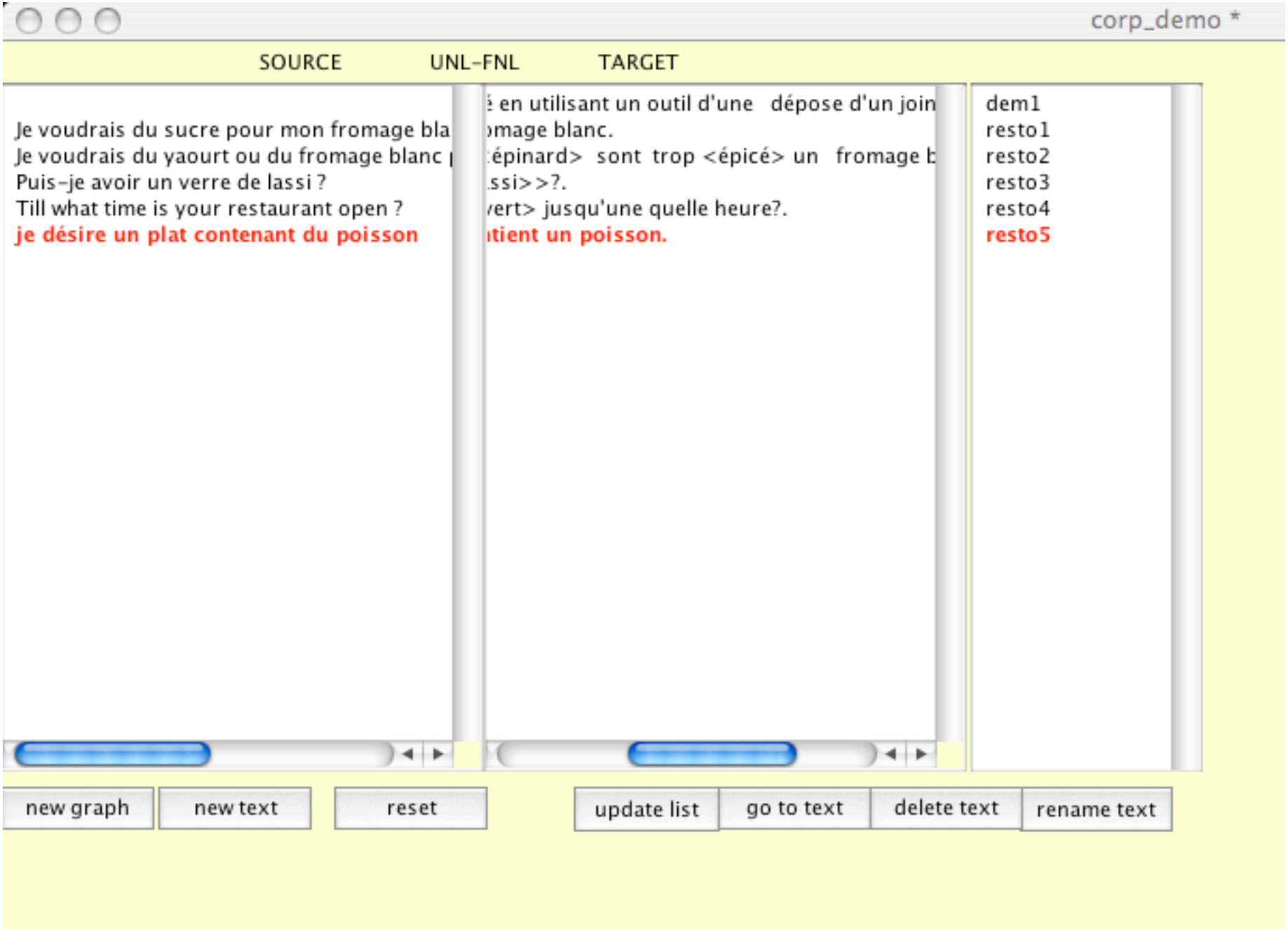
NOTES

<-- | index | --> trace | --> deconverter | --> graphic | -->

May I get a glass of lassi?




corp_demo *
SOURCE
UNL-FNL
TARGET
Je voudrais du sucre pour mon fromage bla
Je voudrais du yaourt ou du fromage blanc
Puis-je avoir un verre de lassi ?
Till what time is your restaurant open ?
je désire un plat contenant du poisson
Le joint défectueux est retiré en utilisant u
Je désire le sucre pour mon fromage blanc.
Je désire <parce_que> les <épinard> sor
Je peux avoir un verre de <<lassi>>?.
Un votre restaurant est'il <ouvert> jusqu'ur
Je désire un <mets> qui contient un pois
dem1
resto1
resto2
resto3
resto4
resto5
new graph
new text
reset
update list
go to text
delete text
rename text


corp_demo *
SOURCE
UNL-FNL
TARGET
Je voudrais du sucre pour mon fromage bla
Je voudrais du yaourt ou du fromage blanc
Puis-je avoir un verre de lassi ?
Till what time is your restaurant open ?
je désire un plat contenant du poisson
é en utilisant un outil d'une dépose d'un join
omage blanc.
épinard> sont trop <épicé> un fromage b
ssi>>?.
vert> jusqu'une quelle heure?.
tient un poisson.
dem1
resto1
resto2
resto3
resto4
resto5
new graph
new text
reset
update list
go to text
delete text
rename text

corp_demo.courant (6) *

resto5

ORG

je désire un plat contenant du poisson

GRAPH

[S]
;<resto5>
obj(want(agt>volitional thing,obj>thing).@entry,dish(icl>food))
aoj(contain(icl>be,obj>thing),dish(icl>food))
obj(contain(icl>be,obj>thing),fish(icl>food))
agt(want(agt>volitional thing,obj>thing).@entry,I(icl>human))
[/S]

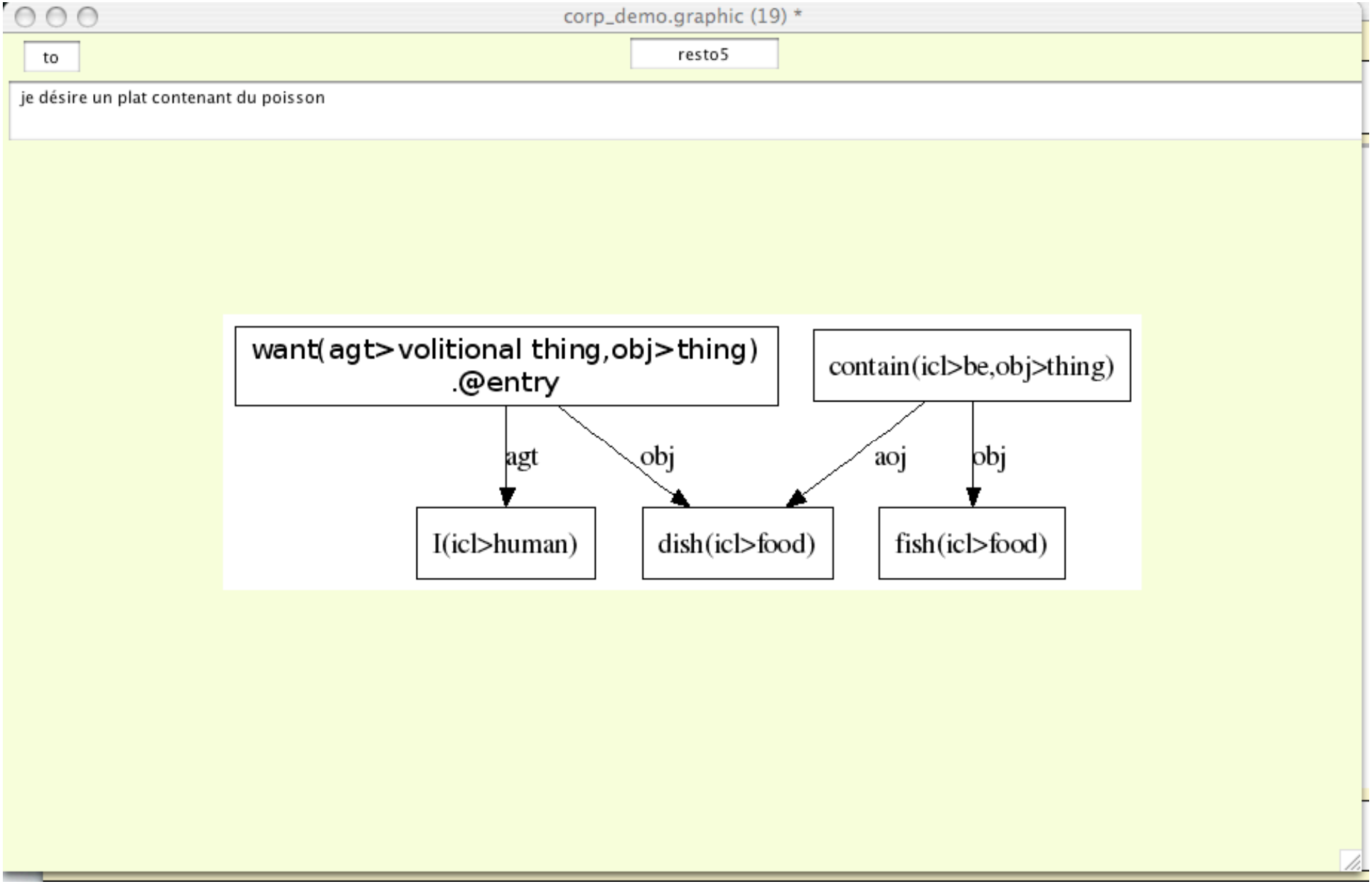
DEC

Je désire un <mets> qui contient un poisson.

NOTES

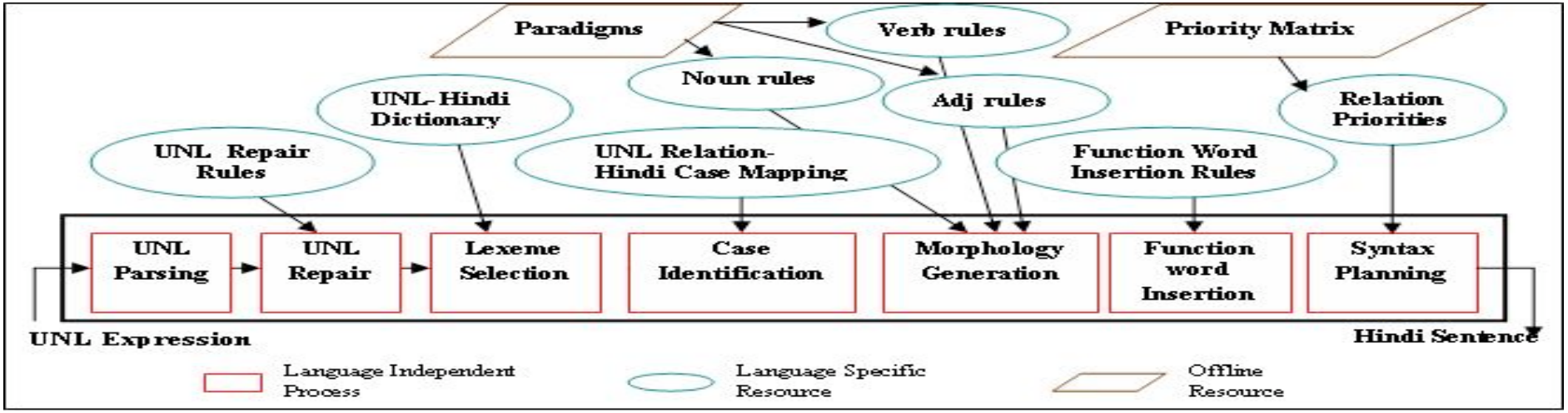
<-- | index | --> trace | --> deconverter | --> graphic | -->

I would like a dish containing fish



Checking UNL graphs by deconversion

- UNL-based "heterogeneous" MT
  - French:
    - unfolding UNL graphs to UNL trees
    - And then to Ariane-G5 trees by lexical transfer
    - Then French generation in Ariane-G5
  - Hindi: following diagram



Results of data construction so far

- Sentences
  - ≈ 450 in F, E, H
    - Very few in Marathi
  - 170 UNL graphs (7/07)
- Vocabulary
  - 350 words
    - Restaurant and Accommodation domains
  - 1000 words targeted for a 1st usable version
- Work will continue 3-4 months in 2008
  - PIVAX should be used
    if synchronization tool ready
    - Various files now: Excel, Revolution, Word, raw text

Conclusion & perspectives

- It seems possible to build useful MT-based applications
  - Using mainly a "table-based" approach yes, Ed!
  - Using "simplified MT" if Noun or Adjective variables
    - deconversion only
    - UNL graphs with variables
  - Using full MT only for unfound sentences
    - If success is only 50%, overall success will be 95% (90+10/2)
    - Call *any* & maybe *several* MT systems
    - Including UNL, Indian systems (between 8 Indian languages) and any F-E, E-I system
- In the context of
  - Academic cooperation
  - Heterogeneous MT system sharing a "lexical pivot"
- Perspective: derive a small F-K-V SurviTra on PDA