A corpus-based NLP-chain for a web-based Assisting Conversational Agent

Mao Xuetao, Jean-Paul Sansonnet, François Bouchet
LIMSI-CNRS

Outline

☐ Problem

- Assisting agents
- Assisting agents for web applications and services
- The genealogy of the DIVA toolkit
- A typical chatbot architecture
- Advantages and drawbacks of the chatbot approach

☐ Methodology

- Methodology: a corpus-based NLP-chain
- The linguistic domain of assisting questions
- Methodology for the corpus collection
- Excerpt from the sub-corpus 'Marco'
- Assistance is a linguistic genre

☐ Implementation

- DIVA NLP-chain
- DIVA semantic keys
- DIVA formalization phase: $\mathfrak{R}$- rules
- DIVA topic files
- DIVA interpretation phase: $\mathfrak{I}$-rules

Conclusion

Can we use the chatbot architectures as a base for the analysis and resolution of natural language assisting requests in web applications and services?

— *Yes, provided we improve drastically their precision and genericity.*

Because the linguistic domain of the Function of Assistance is precise and concise, we can rely on a corpus-based approach to exhibit the inherent generic phenomena.

From the collected corpus we can extract:
— A set of generic formalization rules;
— A set of generic semantic classes;
— A set of generic interpretation rules/classes.

Assisting agents

« An Assisting Agent is a software tool with the capacity to resolve help requests, issuing from novice users, about the static structure and the dynamic functioning of software components or services »

Project InterViews – February 1999
Following Patti Maes MIT, 1994

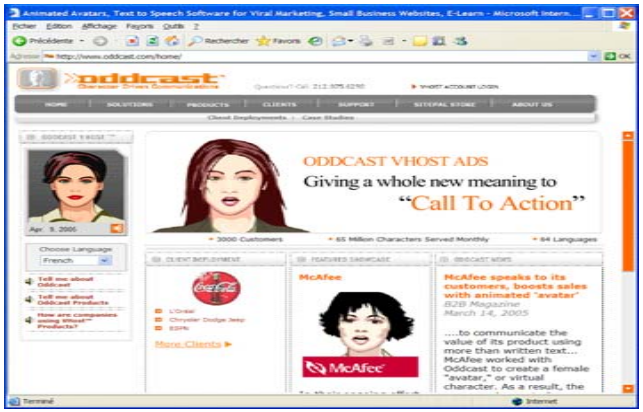
User	person with poor knowledge about the component (novice)
Request	help demand in natural language (speech/text)
Component	computer application, web service, ambient appliance
Agent	rational, assistant, conversational, (can be embodied)
Mediator	symbolic model of the structure and the functioning

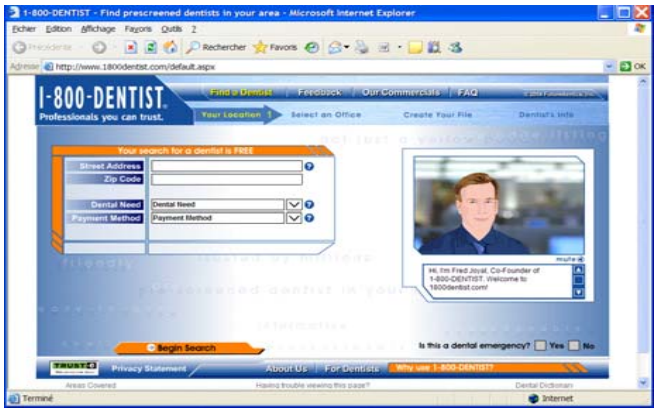
Assisting agents for web applications & services

Keys issues: How can we improve

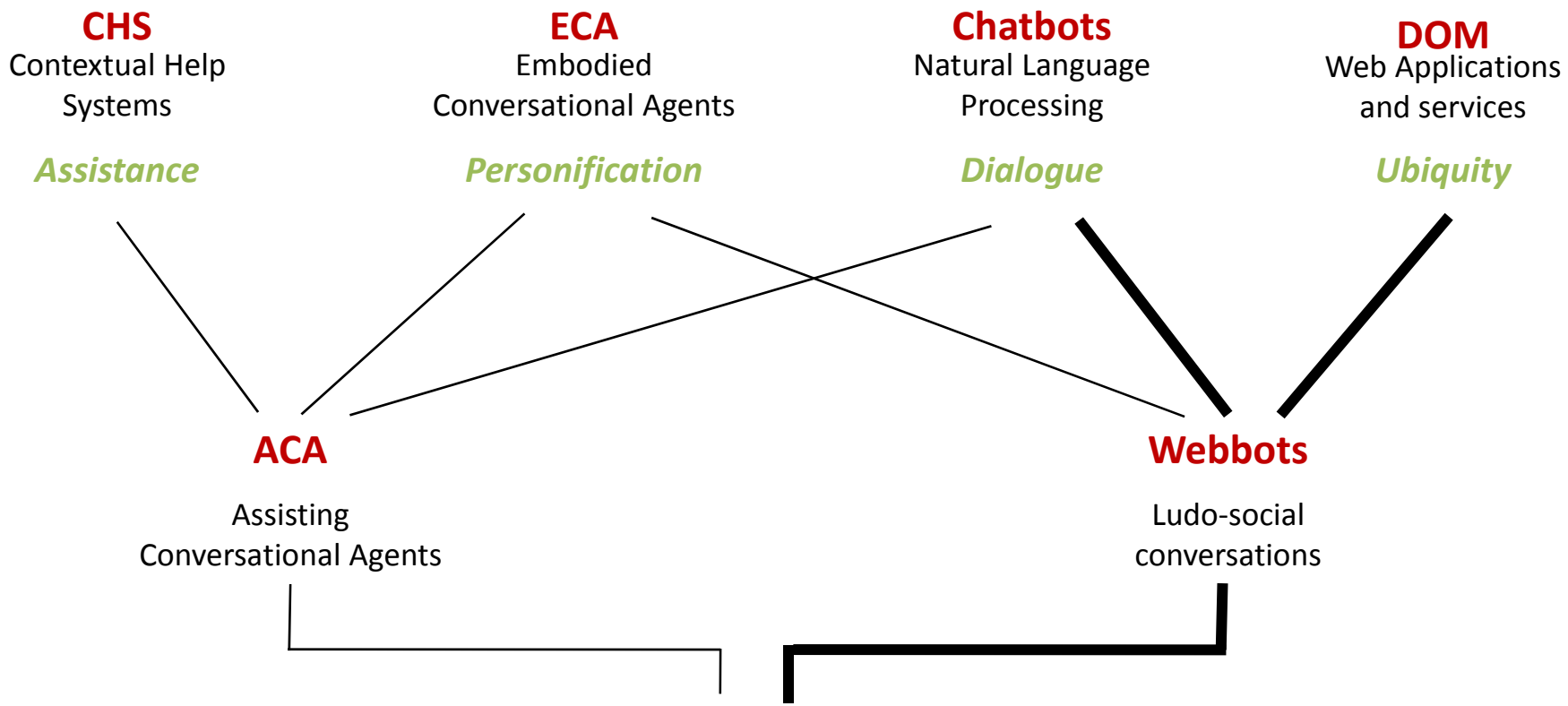
- **The precision**
- **The genericity**

of the Function of Assistance?



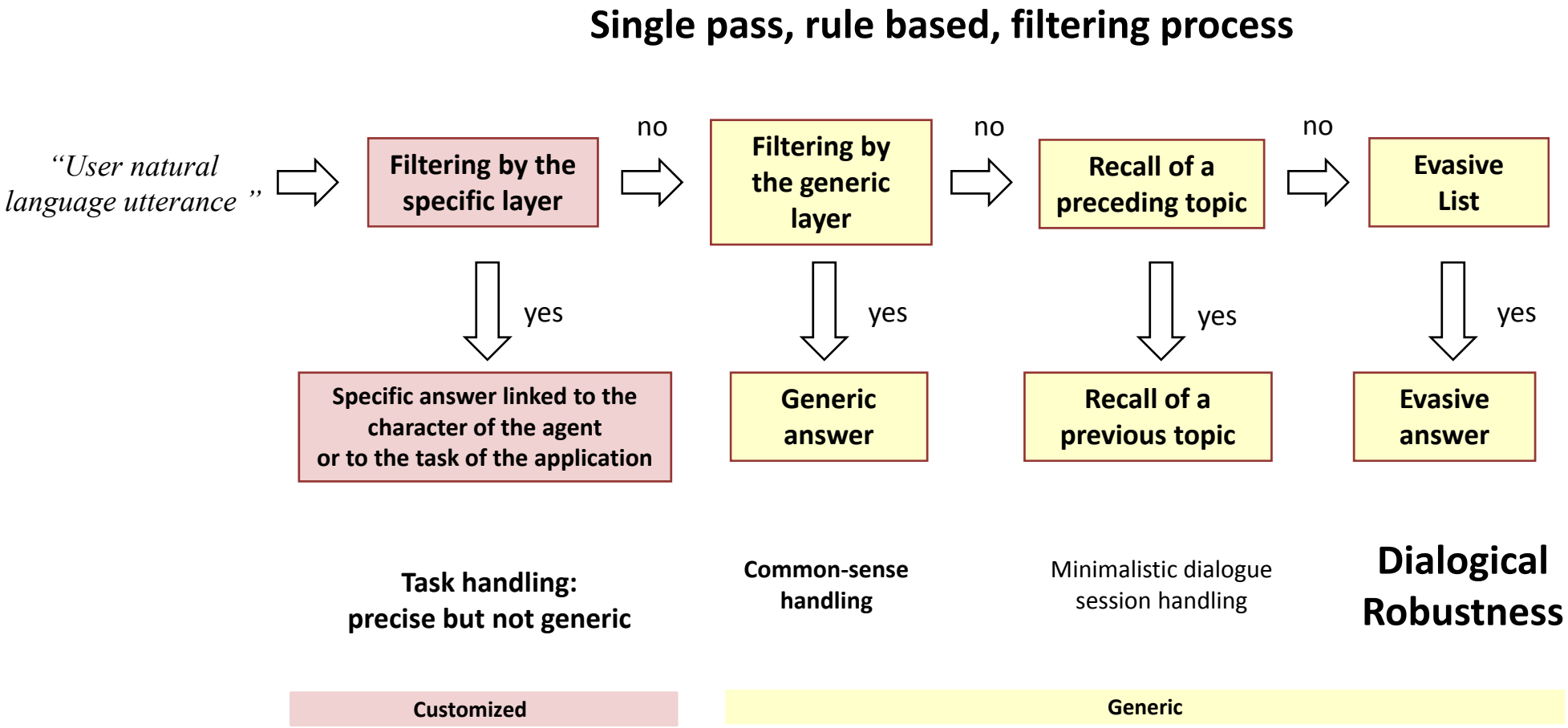


The genealogy of the DIVA toolkit



Problem: How can we improve the precision and the genericity of the NLP-chain in a web environment?

A typical webbot architecture



ALICE's AIML: a simple bot rule

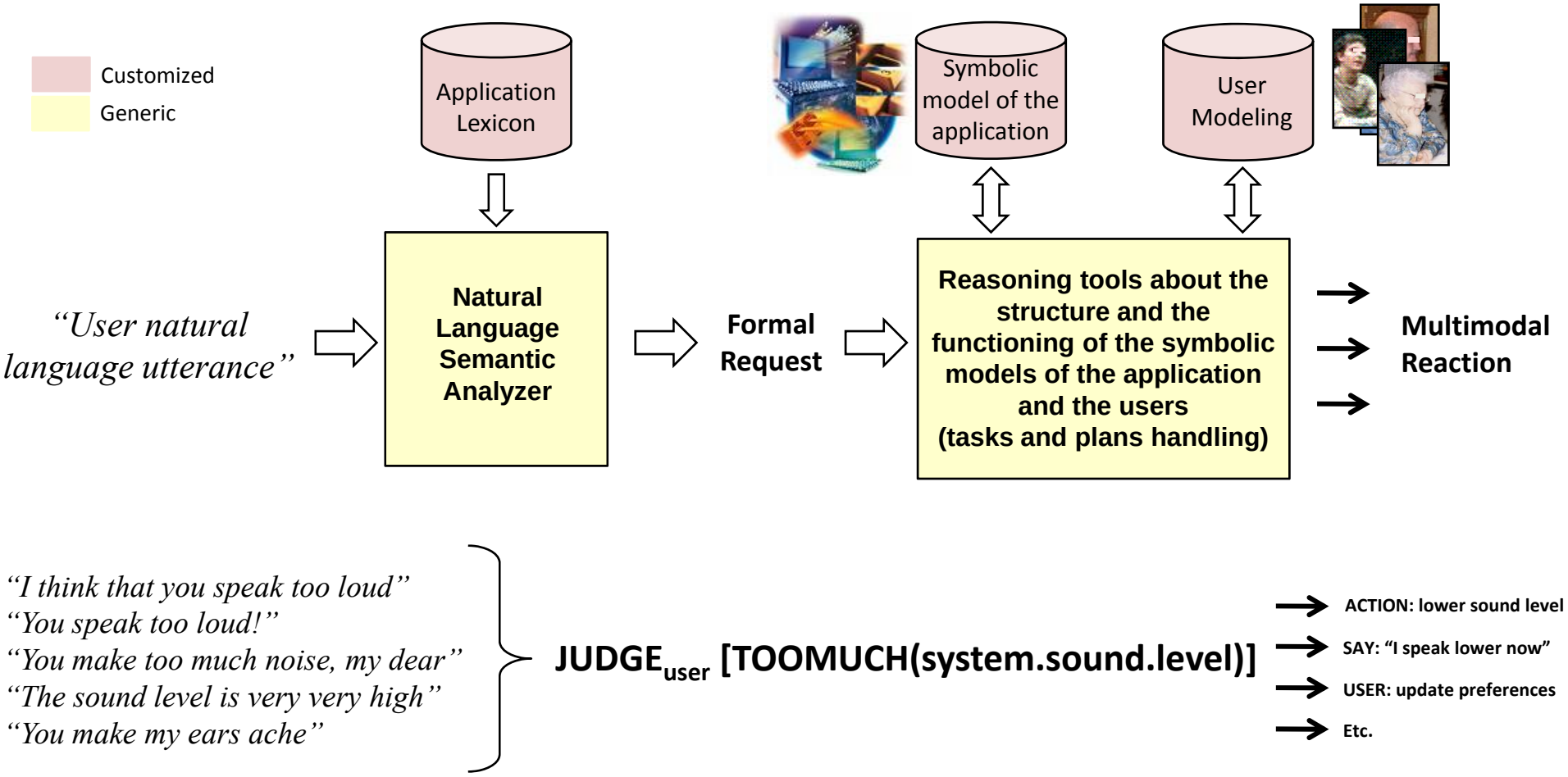
AIML is the format used in Wallace's ALICE chatbot who won several times the *Loebner* prize.
Here is a simple AIML rule (called an atomic category):

```
<category>
        <pattern>WHAT IS A CIRCLE</pattern>
        <template>
        <set_it>a circle</set_it>
         is the set of points equidistant from a common point called the center.
        </template>
</category>
```

The above rule does the following:

1. Matches a user input like this one: *"Can you tell me what is a circle please?"*
2. Sets the internal register "IT" to the value of "a circle" [minimalistic model of the session]
3. Sends the user the answer: *"A circle is the set of points equidistant from a common point called the center."*

A typical finalized dialogue system



Evaluating the linguistic capabilities of chatbots

— Wollermann, C. (2004). Evaluierung der linguistischen Fähigkeiten von Chatbots. Magister report, Rheinische-Friedrich-Wilhelms Universität Bonn.

— Wollermann, C. (2006). Proceedings of the Young Researchers' Roundtable on Spoken Dialogue Systems, 75-76. Pittsburgh, PA, Sept 2006.

"To what extent are chatbot systems able to analyze the users input on the semantic and pragmatic level?"

☐ Evaluation methodology

- Four main chatbots: ALICE, EllaZ, Elbot, ULTRA-HAL-ASSISTANT.
- A collection of linguistic phenomena where evaluated qualitatively in the chatbot answers to users questions:
  - — Semantic: Semantic relations, Quantifiers, Anaphora.
  - — Pragmatic: Grice's maxims.

☐ Results

- Semantic relations: ∅ but for EllaZ which relies on WordNet
- Quantifiers: partly handled, in the four chatbots
- Anaphora: ∅
- Grice's maxims: ∅ (unaccountable in chatbots)

BOTTOM LINE: A deeper semantic/pragmatic analysis is required for finalized/task-oriented dialogue.

QUESTION: Can we improve on the chatbot approach?

Advantages and drawbacks of the chatbot approach

☐ Advantages: easy, light, precise

- They are easy to develop: no large semantic analyzer, no complex reasoning tools;
- They are light to deploy in a web-based environment → client architectures can be envisioned;
- They provide robust natural language reactions (Evasive list effect – ELIZA effect);
- They are tailored and well-suited for the field of ludo-social chat;
- When associated with a given application, they can be customized to be extremely precise.

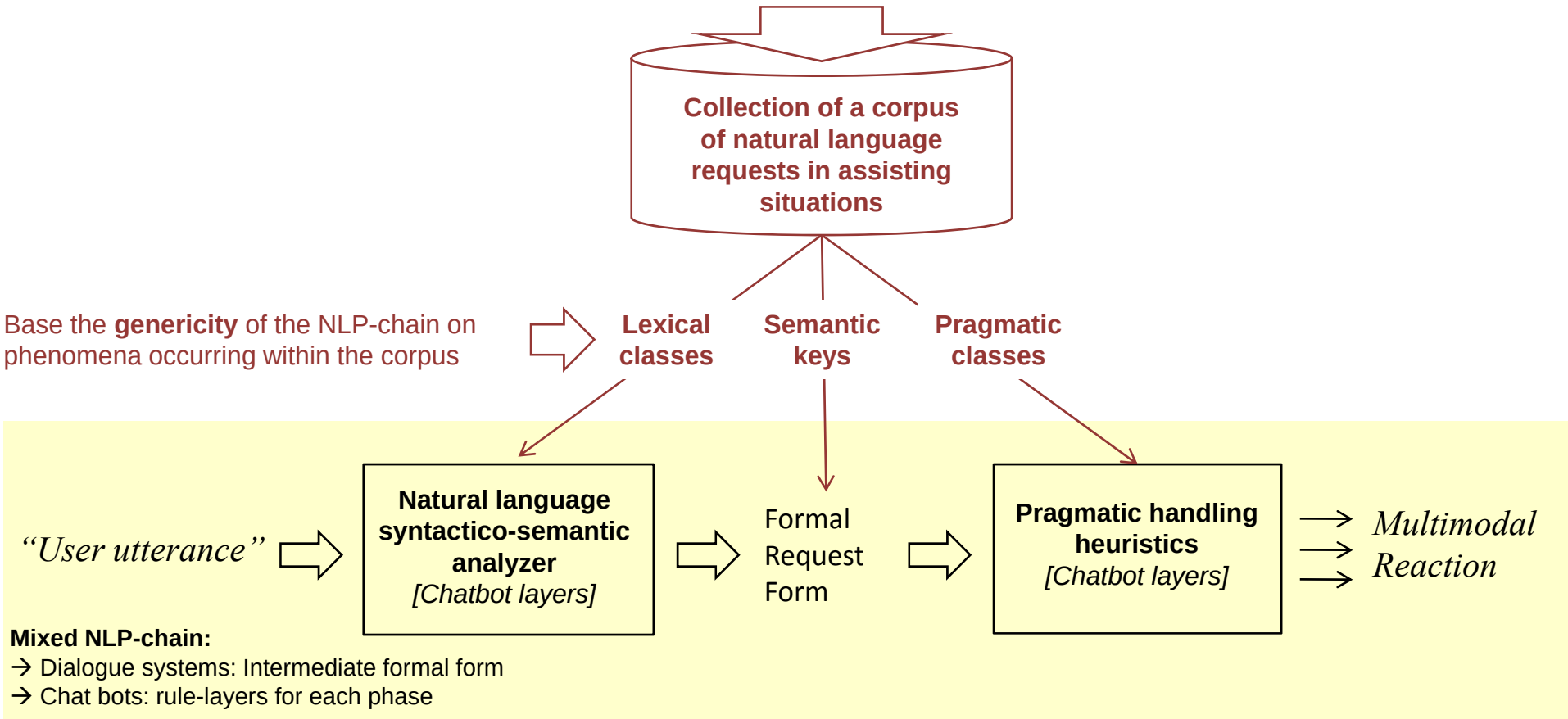
☐ Drawbacks: lack of genericity

- Minimalistic/ultra-customized model of the application;
- Minimalistic model of the dialogue session and of the users;
- No semantic analyzer → lack of precision in the requests (grammar, speech acts, …);
- No formal requests → class reactions are directly linked to specific linguistics patterns;
- No generic reasoning tools, especially when the function of assistance is concerned.

→ Need recoding quite everything for each new application,

→ No reusability, no capitalization.

Methodology: a corpus-based NLP-chain



"If I want to buy such a Scenic, what can I do?" →

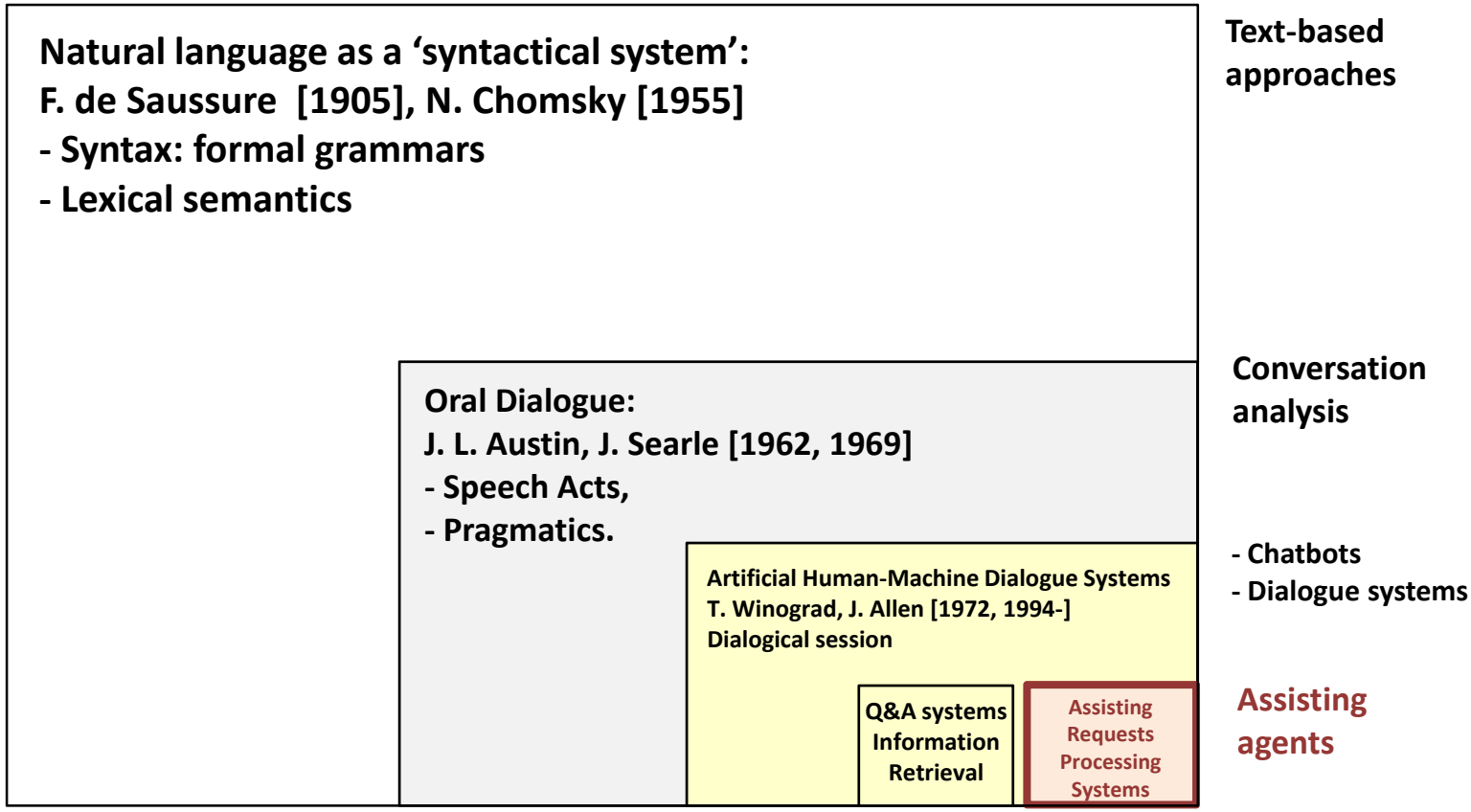
< QUEST IF THEUSER TOWANT TOOBTAIN such a $THECAR WHAT TOCAN THEUSER TODO >

< HOW TOOBTAIN $THECAR>

The linguistic domain of assisting questions

Key hypothesis: The quite restricted linguistic domain concerned makes it tractable:

1. to characterize the distributionality of the linguistic domain,
2. to build a robust semantic analyzer covering the users natural language requests.



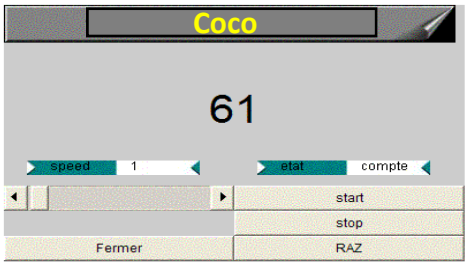
Methodology for collecting a corpus of assistance

☐ Daft 11k corpus content

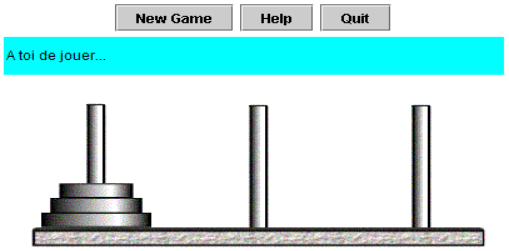
- ~11 000 sentences in French, registered between 2005 and 2007, now continuing…
- Covering: chat activity, control/command activity, direct and indirect assistance requests.

☐ About 2/3 sentences were registered in experimental conditions

- Java Stand alone applications,
- Web sites: LIMSI-AMI, GTACA (corpus Marco online), Webapps of the DIVA toolkit.



Component "Counter"



Component "Hanoi"

Component "AMI web site"

Etc.

☐ About 1/3 sentences were hand-built [for maximizing the coverage]

- From patterns taken from the "Expressways functions" of J. Molinsky et B. Bliss, 1995
- From patterns taken from the "Active Grammar" of the English/French dictionary Robert&Collins – blue pages – B. T. Atkins, M.A. Lewis, D. Feri, H. Bernaert, Ch. Penman. 4th Edition, 1996.

Excerpt from the sub-corpus ‘Marco’ ⊂ Daft 11 k

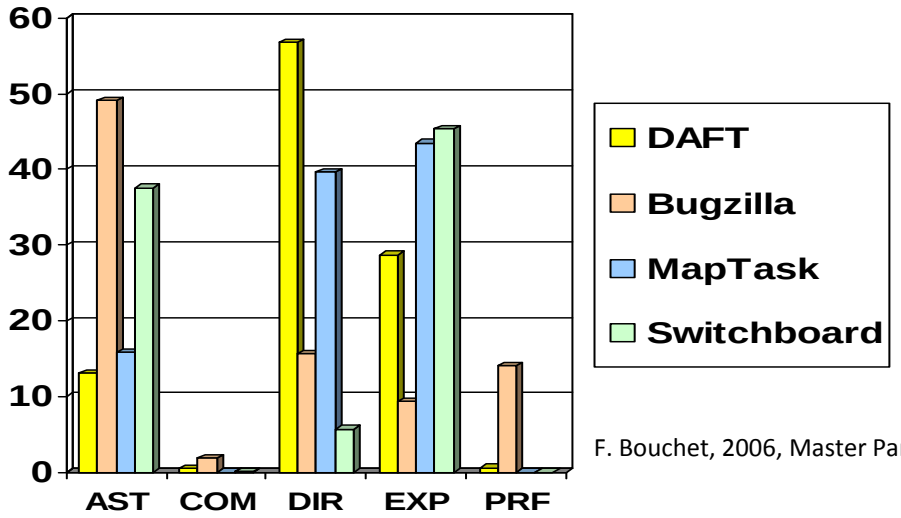
Orthographic noise
Idiosyncratic noise

a+
ah
à l'aide !
Allez à la page des projets
Allez ciao.
Alors ça vient?
alors là t'es completement paumé !
alors lâ t'es completement paumé !
a plus
appelle moi simplement … Sylvie
Appelles moi le manager du site
à quoi penses tu?
A quoi sers-tu?
A quoi sert-tu dis moi un peu?
As-tu des amis?
as tu des idées sur la manière de modifier cette pge ?
as-tu des informations sur les membres du GT ACA ?
as tu des informtion sur comment on peut s’abonner?
as tu entendu parler de Jean-Pierre Durand ?
auf viedersen
au revoir mon vieux
au sujet de cette page, que peux tu dire ?
avec ce corpus, tu sauras ce qu'est une anaphore ...
avec quoi je reviens?
bah!
Bah tu viens de dire que tu pouvais remonter le moral !
barre toi de là
ben alors reponds !!!!!!!!
be ouais tu comprends pas
Bizarre, si je clicke sur le lien du bas ça fait rien
bon
bon â rien !
bon, ça va comme ça !
Bon, dis-moi plutôt ce que tu sais faire plutôt
que de me montrer que tu ne comprends pas
ce que je dis
Bon je me casse. Bye.
bon j'en ai marre je me tire ...
bjr Marco
bonjour, Marco. Qu'est-ce qui te différencie
d'un robot anthropoide?
bonjourmon vieux
bon la on tourne en rond !
bon, reviens à la page d'accueil du site
bon week end
bon y a rien a tirer de toi !!
…

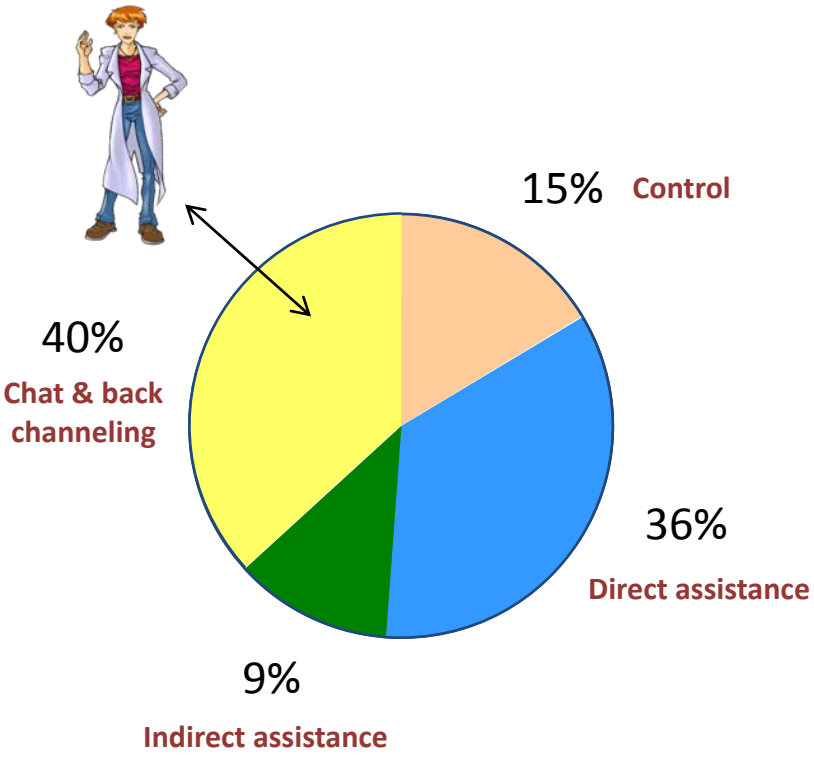
2 sentences

Marco1.0 = 321 utterances with differences at ASCII level

Assistance is a linguistic genre



F. Bouchet, 2006, Master Paris 11



BUGZILLA 6 000 000 comments about correcting Mozilla bugs
MAPTASK 128 dialogues about the building of a geographical map
SWITCHBOARD 200 000 utterances in telephonic conversations

There is a clear "NOT-A-HUMAN" effect:

- More Directives (DIR)
- More Performatives (PRF)
- More Expressives (EXP)
- Less Assertives (AST)
- Lack of Commissives (COM)

DIVA NLP-chain

1. Formalization phase

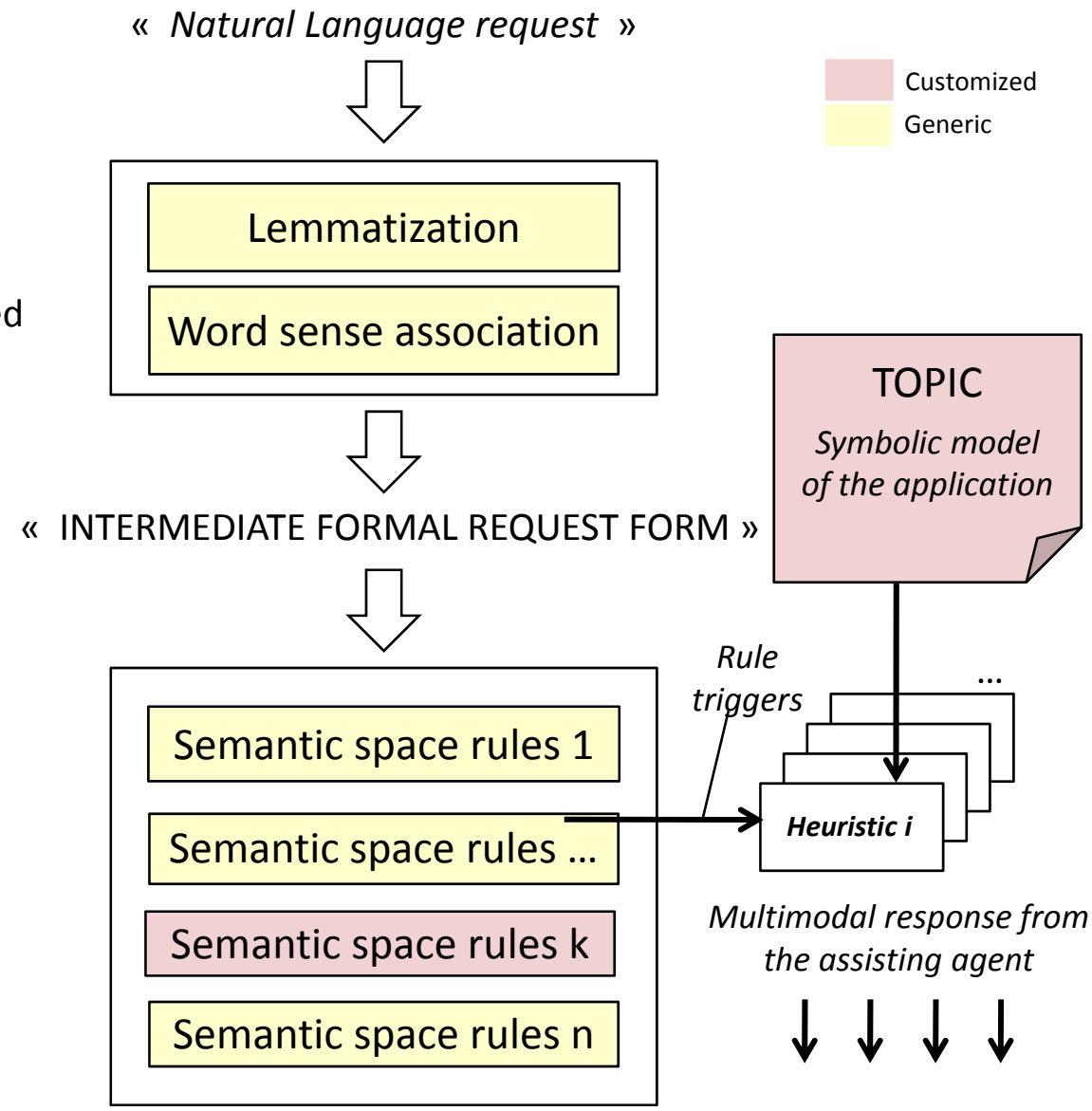
1. Sentences are preprocessed and words are lemmatized;
2. A semantic class (KEY) is associated with each word

2. Interpretation phase

Interpretation rules are of the form:

Pattern → Reaction

Where reactions are expressed as procedural heuristics achieving reasoning tasks over the description of the application (the topic file).



DIVA semantic keys

A “semantic key” is a unique symbol attached to a gloss semantics in a lexicon.

The total number of keys defined from the *manual analysis of the corpus* is 436, divided into six main classes:

NAMES LIST	132
CATEGORIES LIST	20
VERBS LIST	115
ADJECTIVES LIST	60
LOCATIONS LIST	23
GRAMMATICALS & SPEECH ACTS LIST	86

The number of semantic classes was *explicitly* restricted to less than 500 (against >100 000 in WordNet or >20 000 in EuroWordnet).

REASON: the small size of the concerned lexical semantics domain.

PROOF: changing application increases the lexicon by less than 2% with new generic terms.

	Keys	Gloss (as encountered in the analyzed excerpt of the Daft corpus)
Verbs	TOWORK	Denotes the general activity of achieving some work
	TODERIVEFROM	Denotes the abstract action of inheriting/deriving its characteristics from something
	TOKNOW	Denotes the mental action of knowing something
	TOHAVE	Denotes the grammatical auxiliary verb: to have
	TOCAN	Denotes the abstract action of having the general capacity or right of doing something
	TOSAYPLEASE	Denotes the expression of saying please to somebody
	TOSPEAK	Denotes the action of speaking
	TOLIKE	Denotes the mental action of liking/loving something/somebody
	TOWANT	Denotes the mental action of desiring/wanting something or a state of affairs to happen
	TOOBTAIN	Denotes the general action of obtaining/acquiring something or some information
Names	THEAVATAR	Denotes the graphical/dialogical assisting character of the application
	THEHELP	Denotes the service/help provided by somebody
	THEMAXIMUM	Denotes the maximum value that a variable can take
	THEUSER	Denotes the user of the application at first person: I, me, myself
	THETITLE	Denotes the title of a window or a frame in the window of the application
	THEPICTURE	Denotes a picture in the window of the application
	THENUMBER	Denotes the count of something/persons
Adjectives	ISHONEST	Denotes the quality of somebody who is honest/sincere
	ISFEMALE	Denotes the quality of a person with gender: female
	ISREAL	Denotes the quality of something that is real/physical
	ISSAME	Denotes the quality of something that is equivalent/identical/similar to something
	ISUNPLEASANT	Denotes the quality of something that is unpleasant
	ISUNFRIENDLY	Denotes the quality of being unfriendly/impolite with somebody
	ISMANDATORY	Denotes the quality of something that is legally/physically mandatory/indispensable
Grammaticals	WHAT	Denotes the grammatical WH-pronoun: what
	WHY	Denotes the grammatical relation: why
	WHERE	Denotes the WH-question: asking for the location of something
	NEG	Denotes the grammatical relation: negation
	QUEST	Denotes the grammatical relation: question
	UNDEFPRON	Denotes the grammatical pronoun: one
	LESSTHAN	Denotes the quality of something that is less than another thing =!= ISLOWERTHAN
	IT	Denotes the grammatical pronoun: it
	TOBE	Denotes the grammatical auxiliary verb: to be

DIVA formalization phase: ℜ- rules

☐ Formalization phase: ℜ- rules

- Syntax: only the pat attribute is mandatory. W_i are chunks matched and extracted by the JavaScript RegularExpression (the order of Wi can be changed in the output).

<rule id = ”ruleid”
pat = ”JavaScript RegularExpression”
if = ”boolean condition guarding the pattern matching”
go = ”continuation to the next rule” >
<filter>[w_1,w_2, .. w_n]</filter>
</rule>

Example 1: a ℜ-rule catching a grammatical form like a negative phrase:

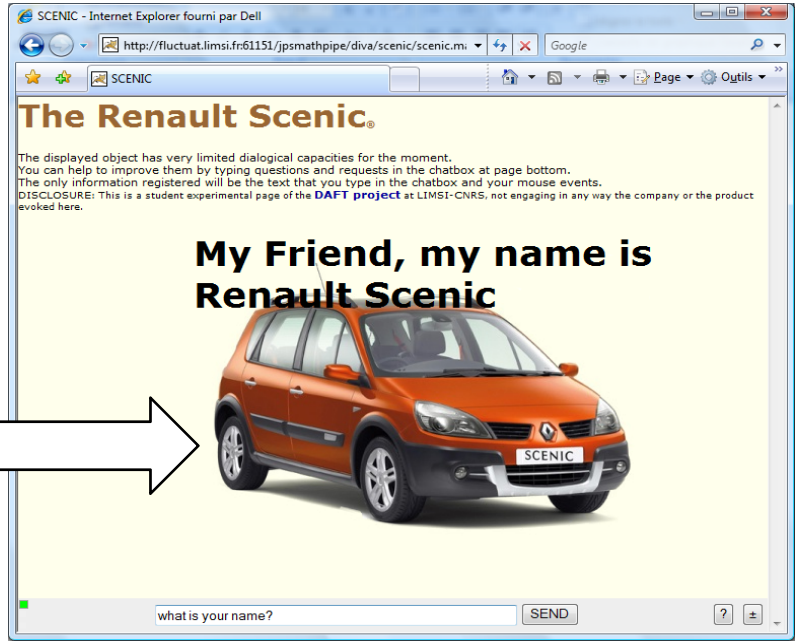
```
<rule id="neg1"
   pat="&lt;(.*)( am | are | is | were )not (.*)&gt;"
   go="NEXTRULE">
   <filter>["NEG","BE",1,3]</filter>
</rule>
```

Example 2: a ℜ-rule catching various flexions associated with the concept ISSIMPLE:

```
<rule id="lem332"
  pat="&lt;(.*)(easy|straightforward|uncomplicated
     |trouble (?: )?free|undemanding|effortless) (.*)&gt;"
 go="NEXTRULE">
 <filter>[1,"ISSIMPLE",3]</filter>
</rule>
```

DIVA topic files

```
<xml>
…
<topicname>TOPICSCENIC</topicname>
<objName>Renault Scénic</objName>
<objAlias encoding="JS">["Scénic"]</objAlias>
<objType>car</objType>
<objSubType usermodify="edit">compact MPV</objSubType>
<objBriefIntro usermodify="edit">The Renault Scénic is a compact MPV
produced by French automaker Renault the first to be
labelled as such in Europe.It is based on the chassis of the Mégane small family car.
It became European Car of the Year on its launch in late 1996.</objBriefIntro>
<objSize>small</objSize>
<objLength encoding="JS" unit="m">4.1</objLength>
<objWidth encoding="JS" unit="m">2.0</objWidth>
<objHeight encoding="JS" unit="m">1.5</objHeight>
<objDiameter encoding="JS" unit="m">null</objDiameter>
<objWeight encoding="JS" unit="kg">2205</objWeight>
<objMaterial>mainly steel</objMaterial>
<objShape>car</objShape>
<objColor usermodify="edit">red</objColor>
<objSmell usermodify="edit"></objSmell>
<objTaste usermodify="edit"></objTaste>
<objTouch>machinery</objTouch>
<objSound encoding="JS">null</objSound>
<objUseHow usermodify="edit">trigger it and drive it</objUseHow>
<objUseRequires encoding="JS" usermodify="edit">["gasoline""some water""road""driver"]</objUseRequires>
<objInputs encoding="JS" usermodify="edit">["gasoline""some water"]</objInputs>
<objOutputs encoding="JS">["Dynamic power""electric power"]</objOutputs>
<objCondition>intact</objCondition>
<objState>idle</objState>
<objAnalogs encoding="JS" usermodify="edit">["Toyota xxx""Audi xx"]</objAnalogs>
</xml>
```



The topic is an XML file containing the description of the static and the dynamic information about a typical 'domain of interest' that is presented to the users on a DIVA web page.

DIVA interpretation phase: $\mathfrak{I}$-rules

Interpretation phase: $\mathfrak{I}$-rules

- Syntax: same as $\mathfrak{R}$-rules with <filter> replaced by one of the following actions (each of them coded in JavaScript):

<do> executes an action on the DOM structure of the page;
<say> makes the agent display a textual answer in its balloon;
<saylater> idem to <say> but the answer is delayed;
<hint> displays a help message in the chatbox bar.

Example

- Suppose the user gives her name with the utterance: "*My name is Jane*"
- The formalization phase can produce the formal request: "USERNAME TOBE jane"

```
<rule id="name2"  pat="&lt; USERNAME BE (\w+) &gt;" >
   <do>
    THETOPIC.x = TALK_capitalizefirst(TALK_getmatch(1));
    If (THETOPIC.x == THEUSER.name)   TALK_say(['I knew it already', 'You said it'] ,0, 2);
    else THEUSER.name = THETOPIC.x;
  </do>
  <say>
   <p>From now I will call you _THETOPIC.name_.</p>
   <p>Ok you name is _THETOPIC.name_ ...</p>
   <p>Ok you are _THETOPIC.name_"</p>
   <p>OK for calling you @</p>
  </say>
</rule>
```

THETOPIC.x ← "Jane"

THEUSER.name ← "Jane"

In topic file

"*From now I will call you Jane*"

Conclusion

Key issues

- Can we develop a cost-effective, web-based, Assisting Conversational Agent?
- How can we improve the precision and the genericity of the traditional chatbot NLP-chain architectures?

Methodology

- Characterize the concerned linguistics domain through the collection of a corpus of questions
- Propose a mixed-approach NLP-chain based on:
  1) An intermediate formal form → base the generic semantic classes on the corpus
  2) Chatbot rule layers for each phase → base the generic pragmatic classes on the corpus

Results

- The DIVA toolkit is operational and available as a support for teaching and research purposes
- The DIVA corpus-based NLP-chain is operational for English [Xuetao, 2008]
- Presently, 24 web applications have been implemented in DIVA: http://www.limsi.fr/~jps/online/diva/divahome

Perspectives

- Propose corpus-based NLP-chain for French
- Merge the resources of the DIVA toolkit (Keys, Rules, XML-files) as a subset of the GRASP-DAFT project.