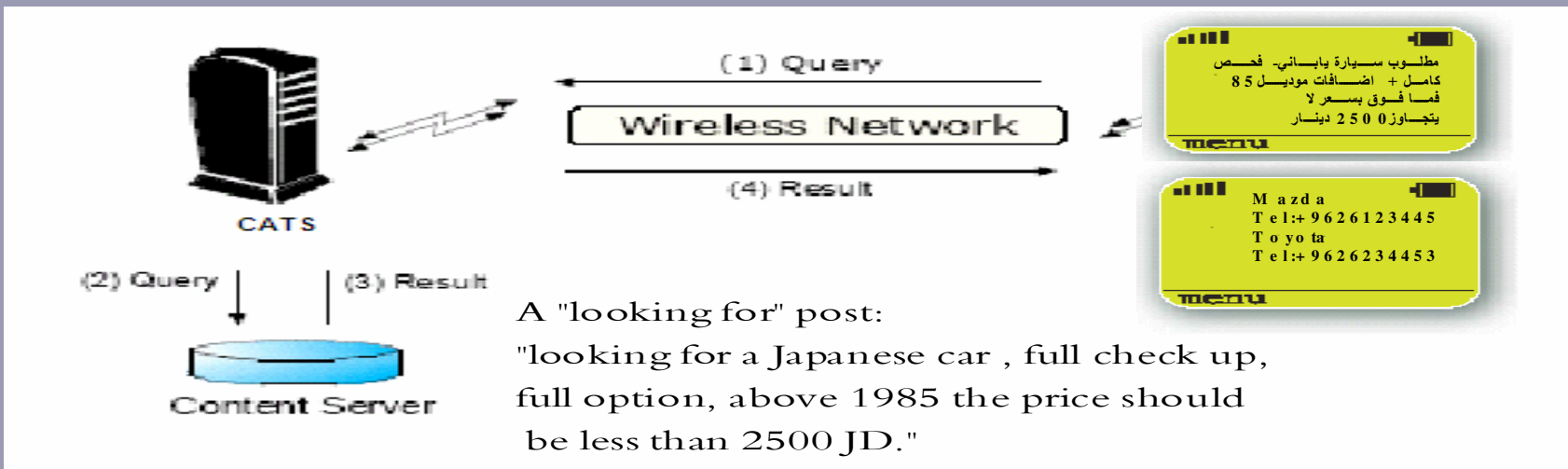
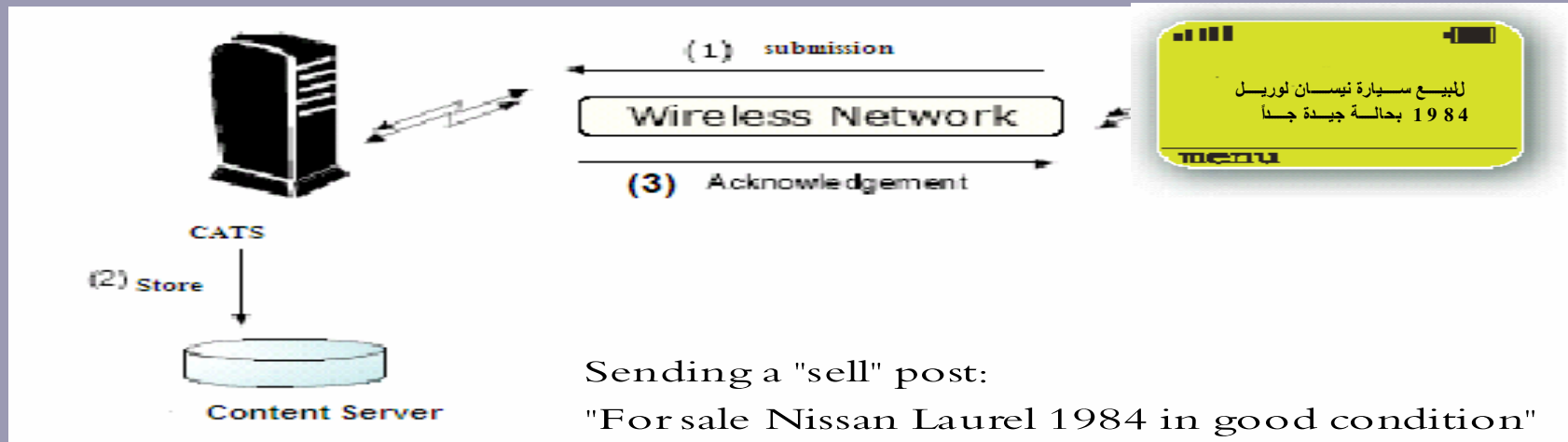


**METHODS FOR HANDLING  
SPONTANEOUS E-COMMERCE  
ARABIC SMS: CATS, AN  
OPERATIONAL PROOF OF CONCEPT**

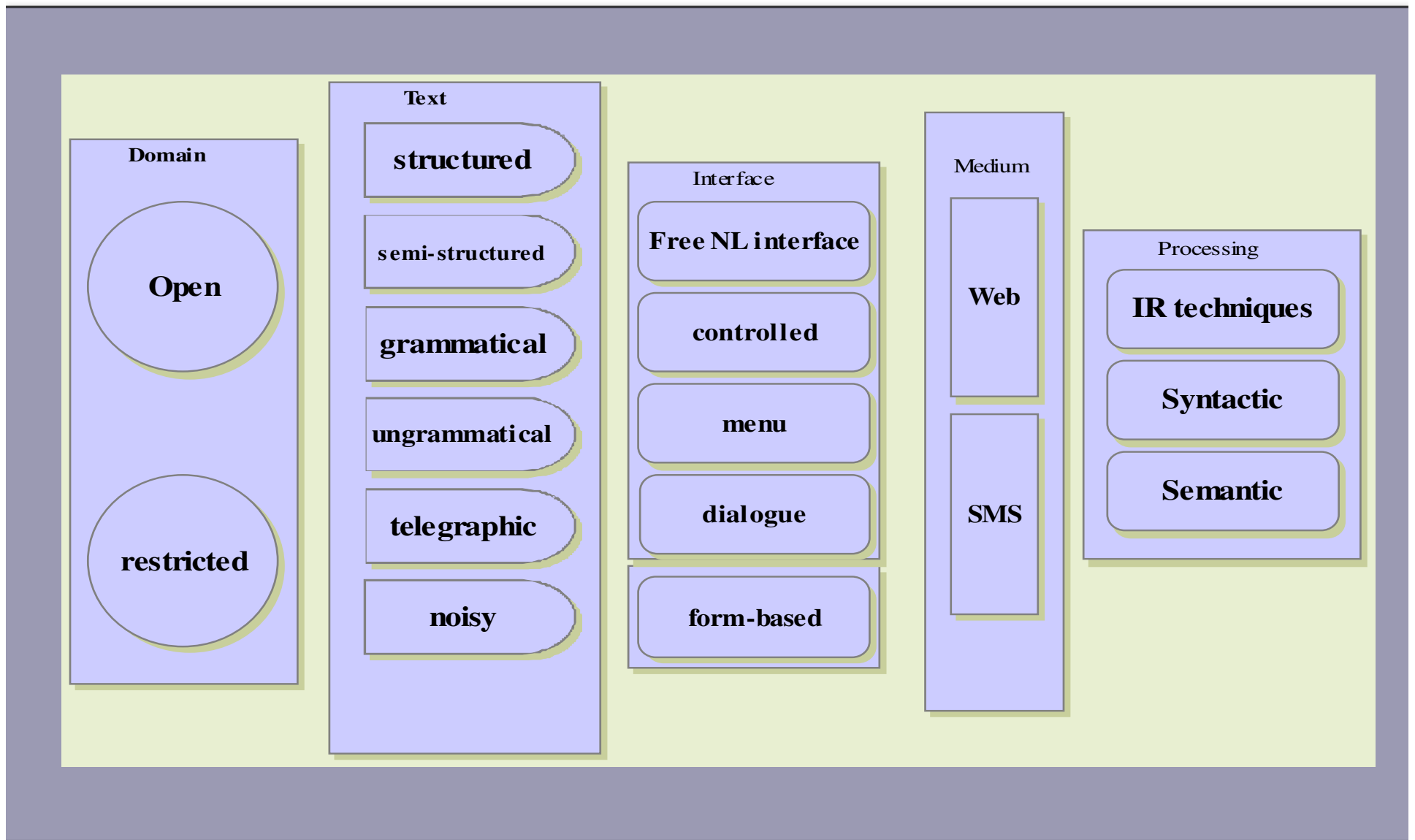
Daoud Maher Daoud



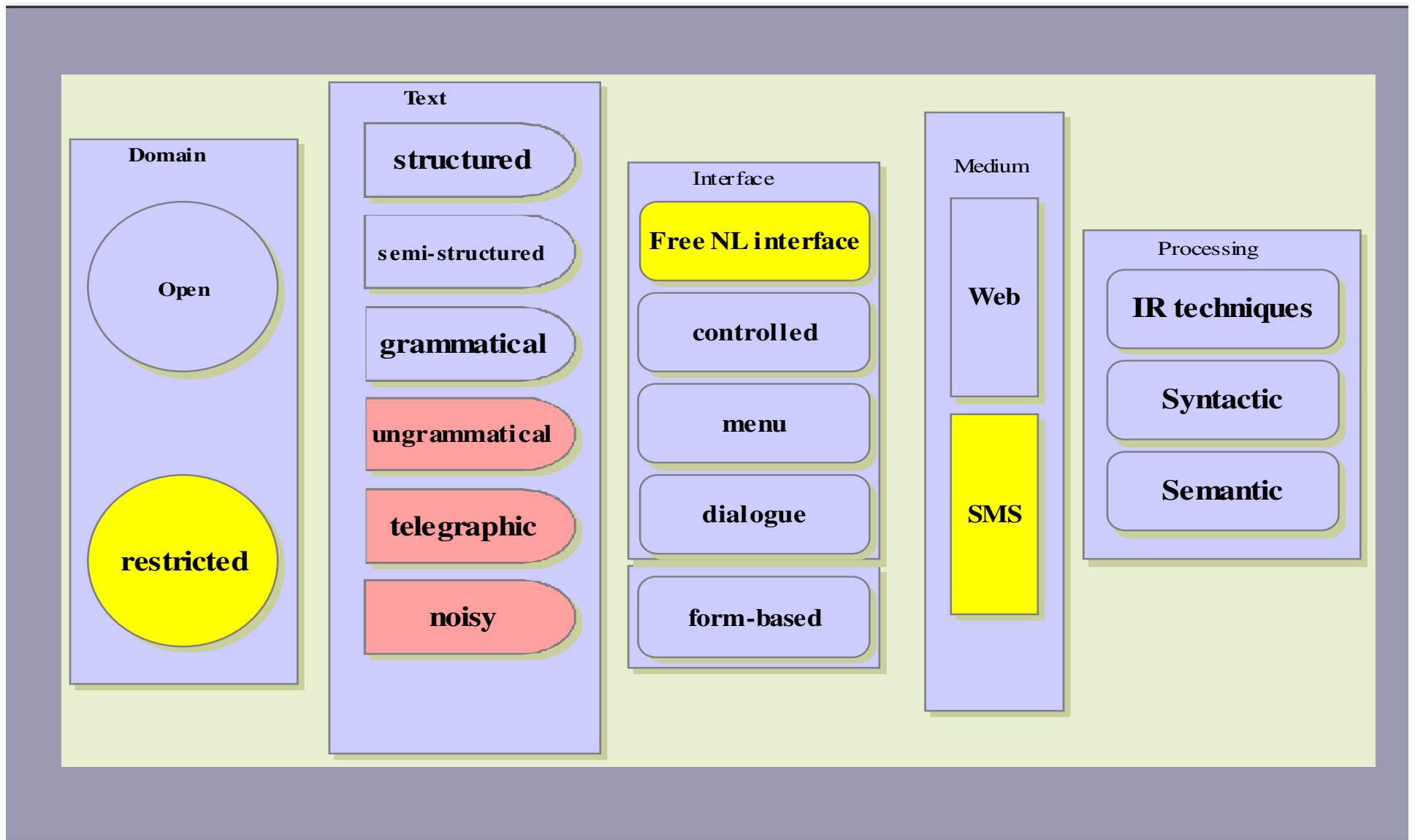
# Idea of a new service: connect buyers and sellers



# Important factors in building such systems

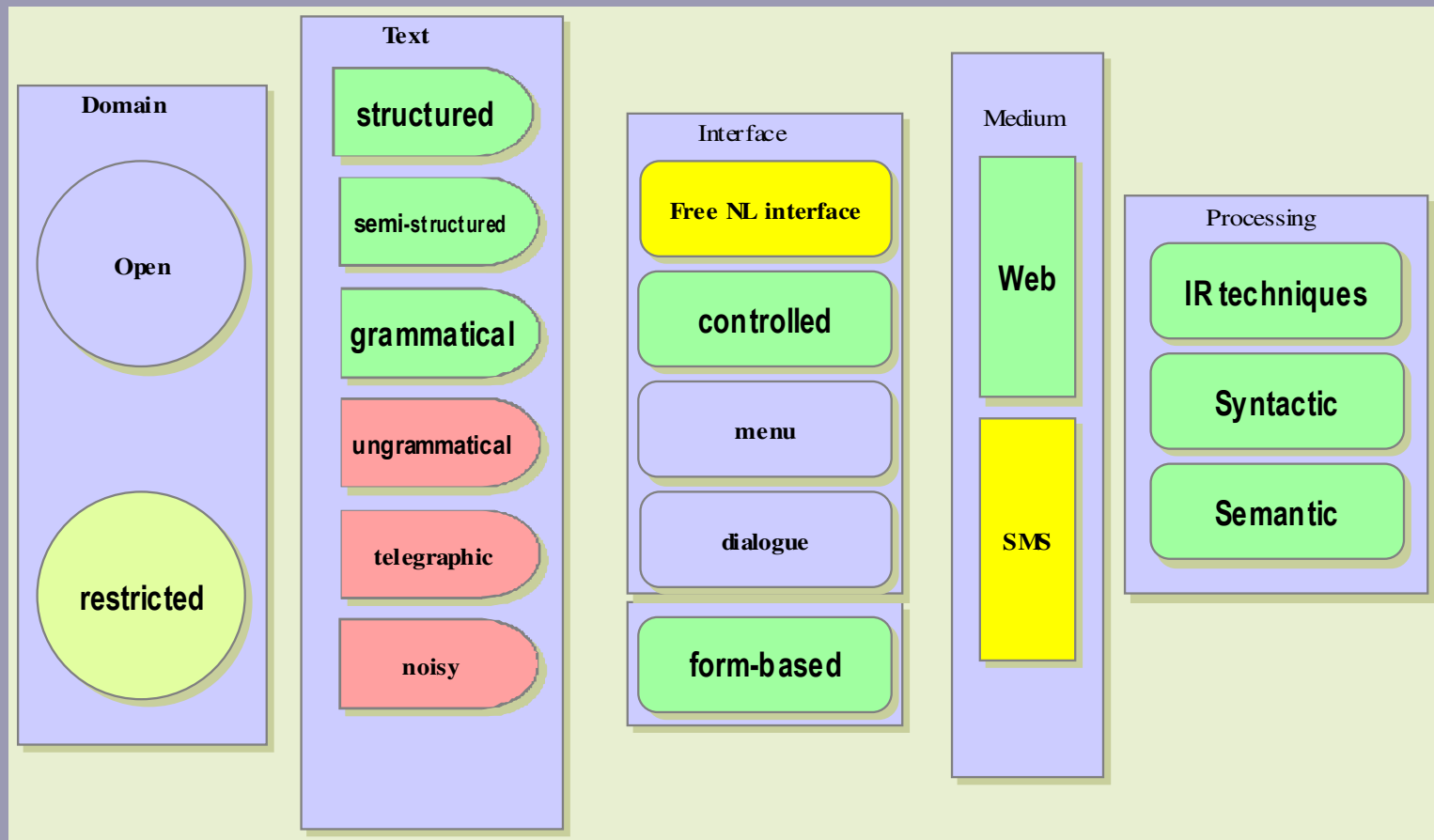


# Necessary choices for the envisaged system

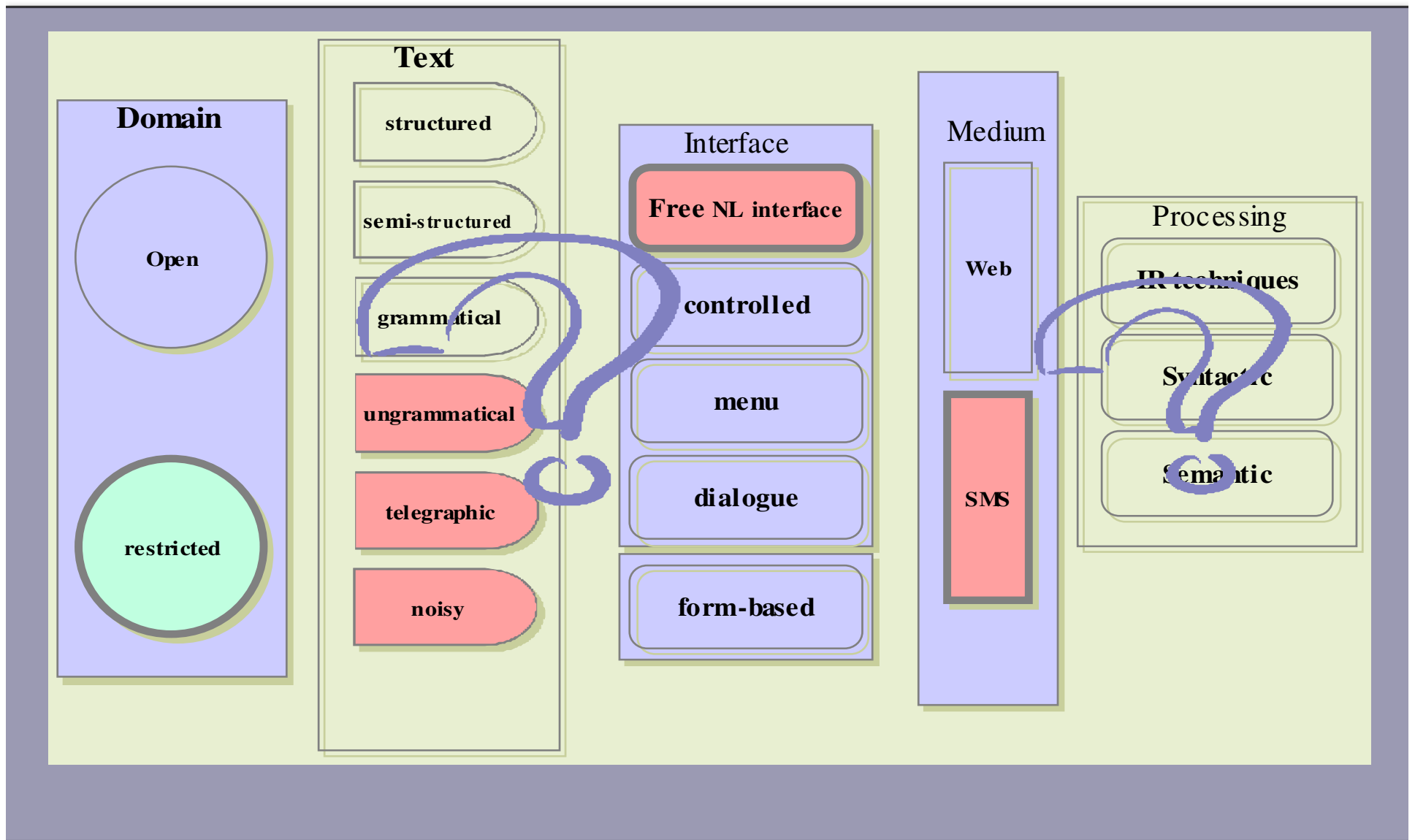


# Current Scene: studied e-commerce systems

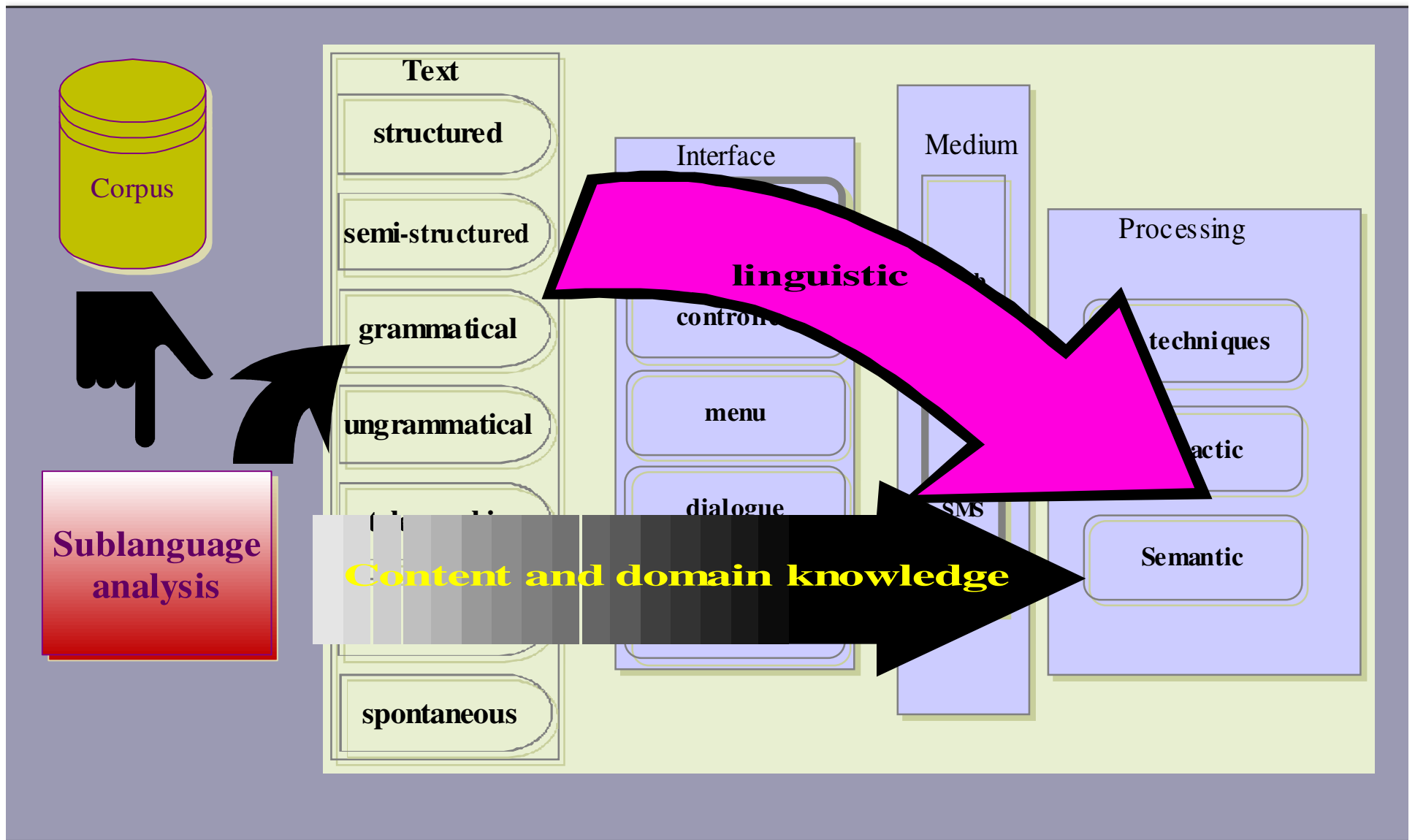
CASA, TREE, MKBEEM, HappyAssistant, MIETTA, Google SMS



# Current Scene: needs not covered

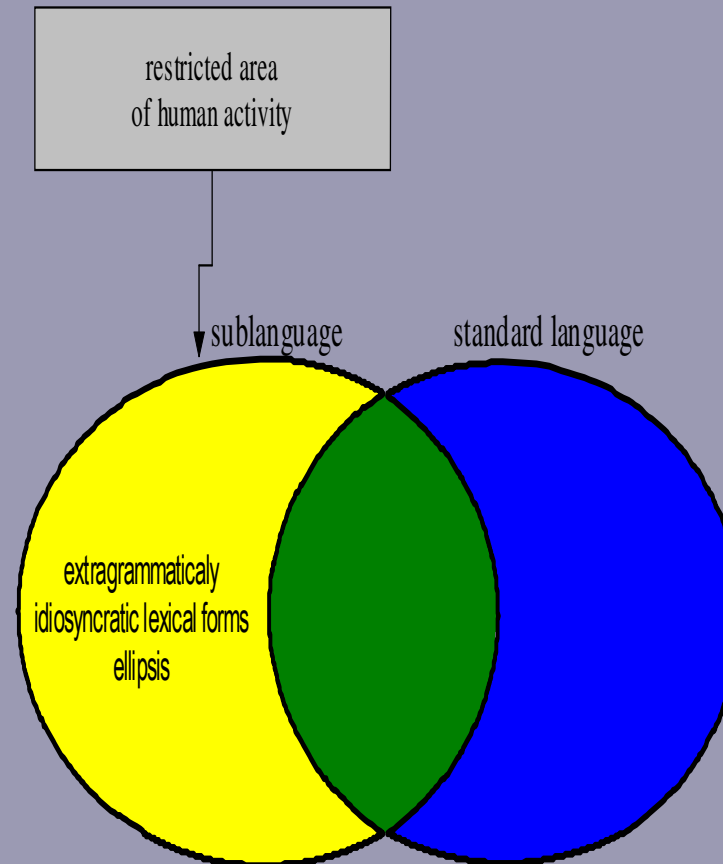


# Current Scene: hypothesis



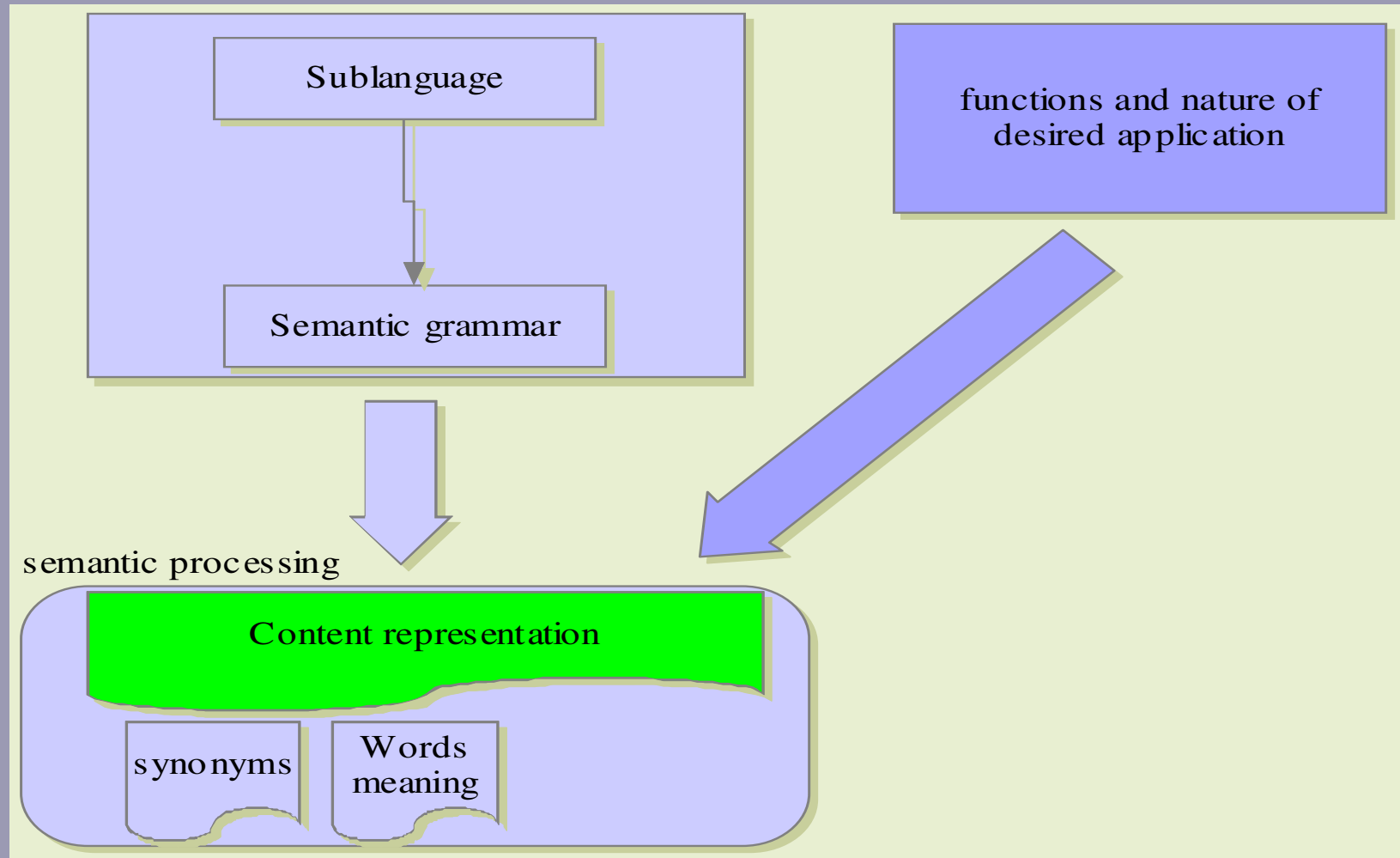
# Impossibility of using “general language techniques”

Several efficient applications have been developed by exploiting the restrictions of a sublanguage (Sager 1986; Kittredge and Lehrberger 1982; Grishman and Kittredge 1986). (LSP, PROTEUS, PUNDIT, KERNEL, and MedLEE, GENIES)

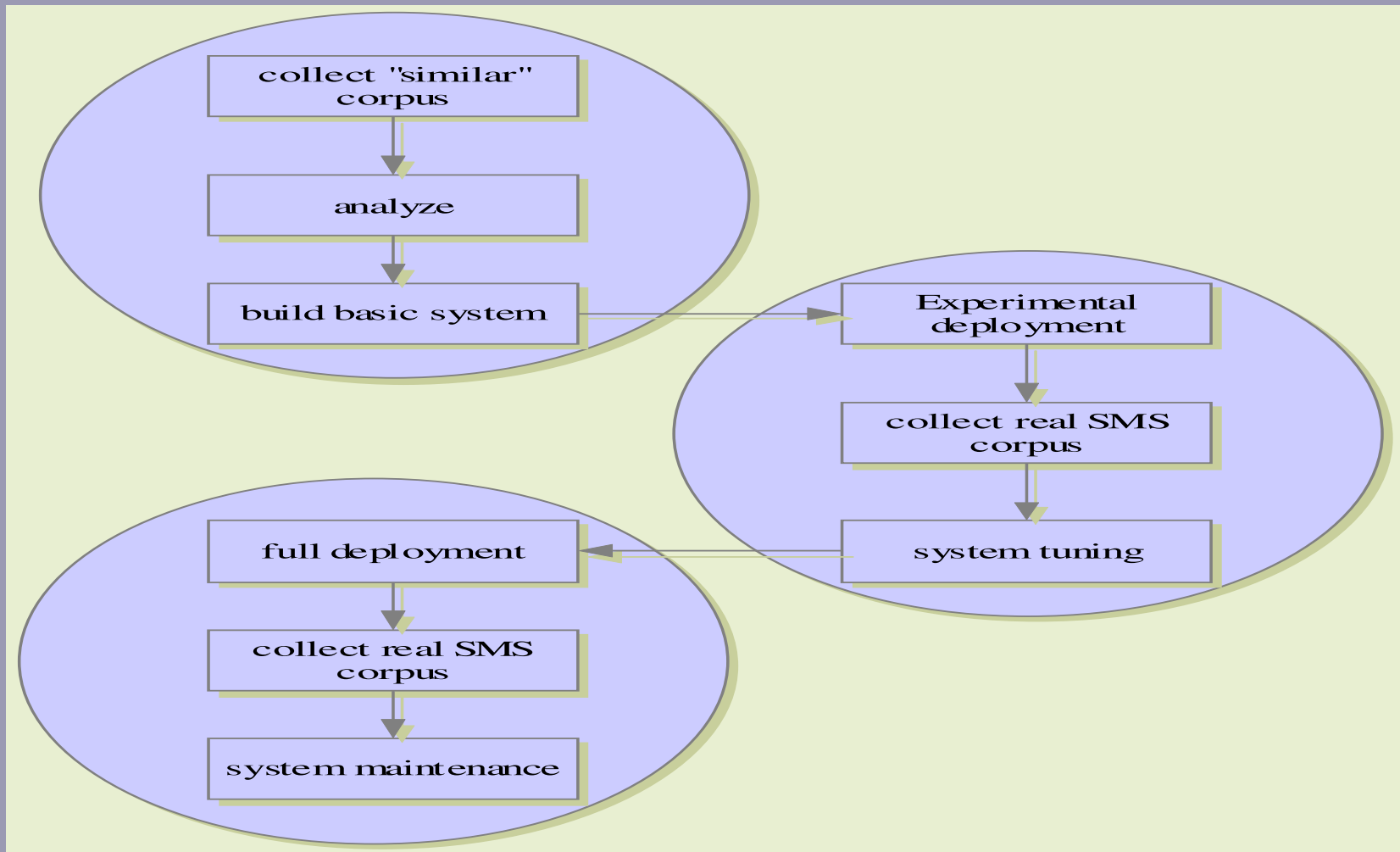


The presence in some sublanguages of structures which are unknown in the standard language . . . no single parsing grammar will be adequate for all types of text.

# Importance of using content representation



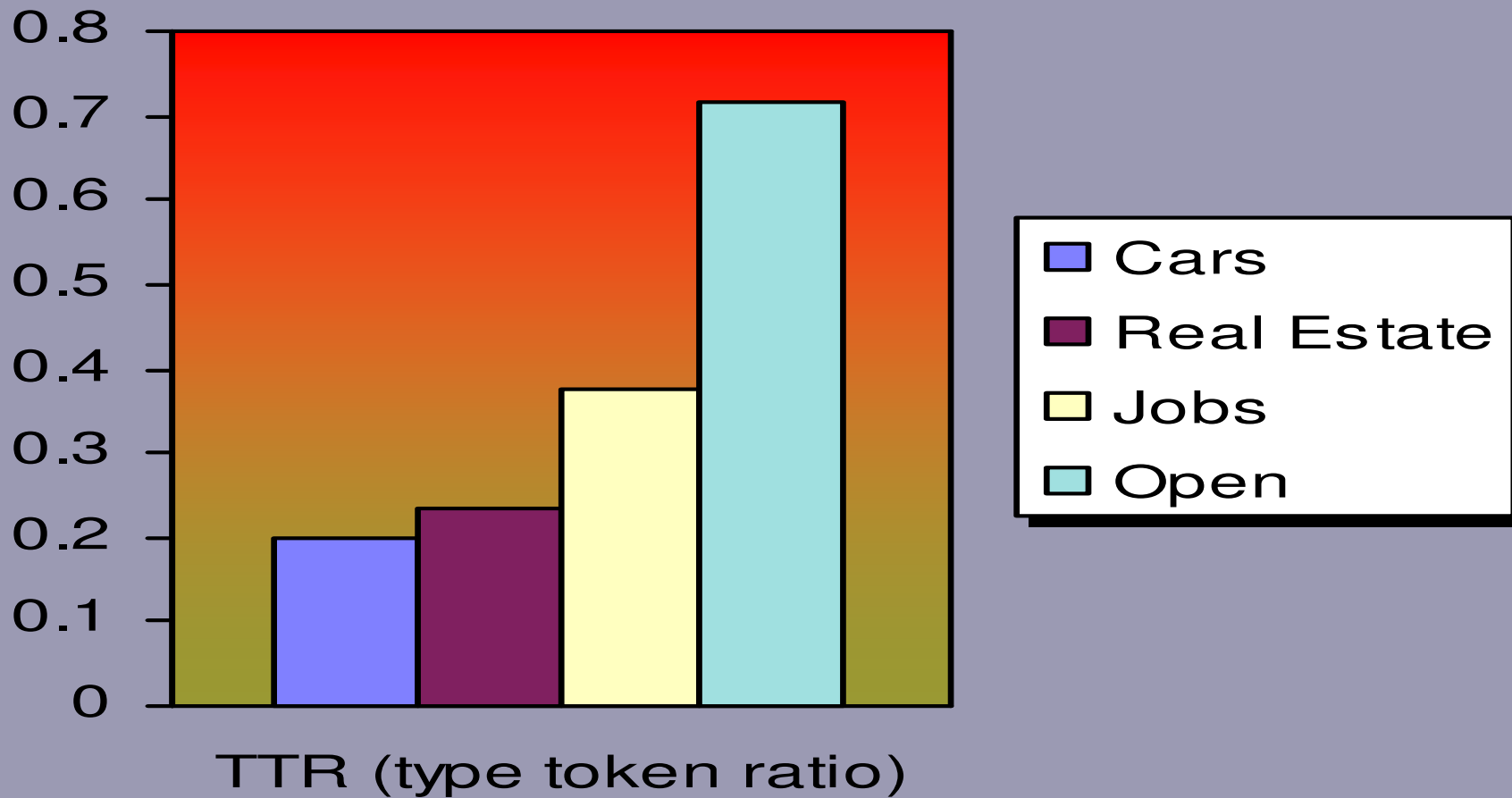
# Sublanguage: the scarcity of data



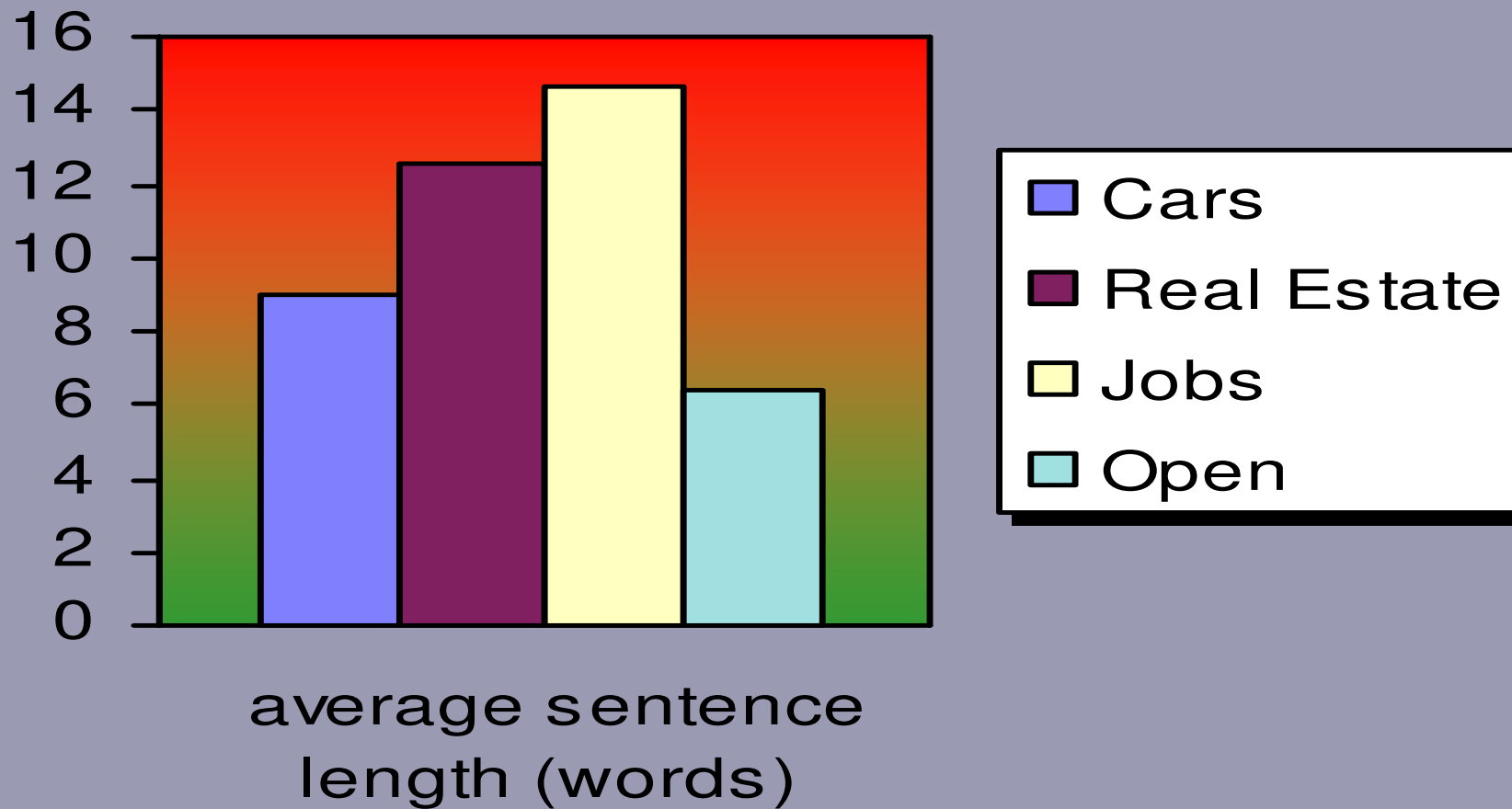
# Sublanguage: analysis approach

- Quantitative properties:
  - Type-Token Ratio (TTR): lexical complexity
  - Sentence length: language complexity
  - Words frequency: identifies the nature of text (telegraphic or normal)
- The manual study for lexico-semantic patterns

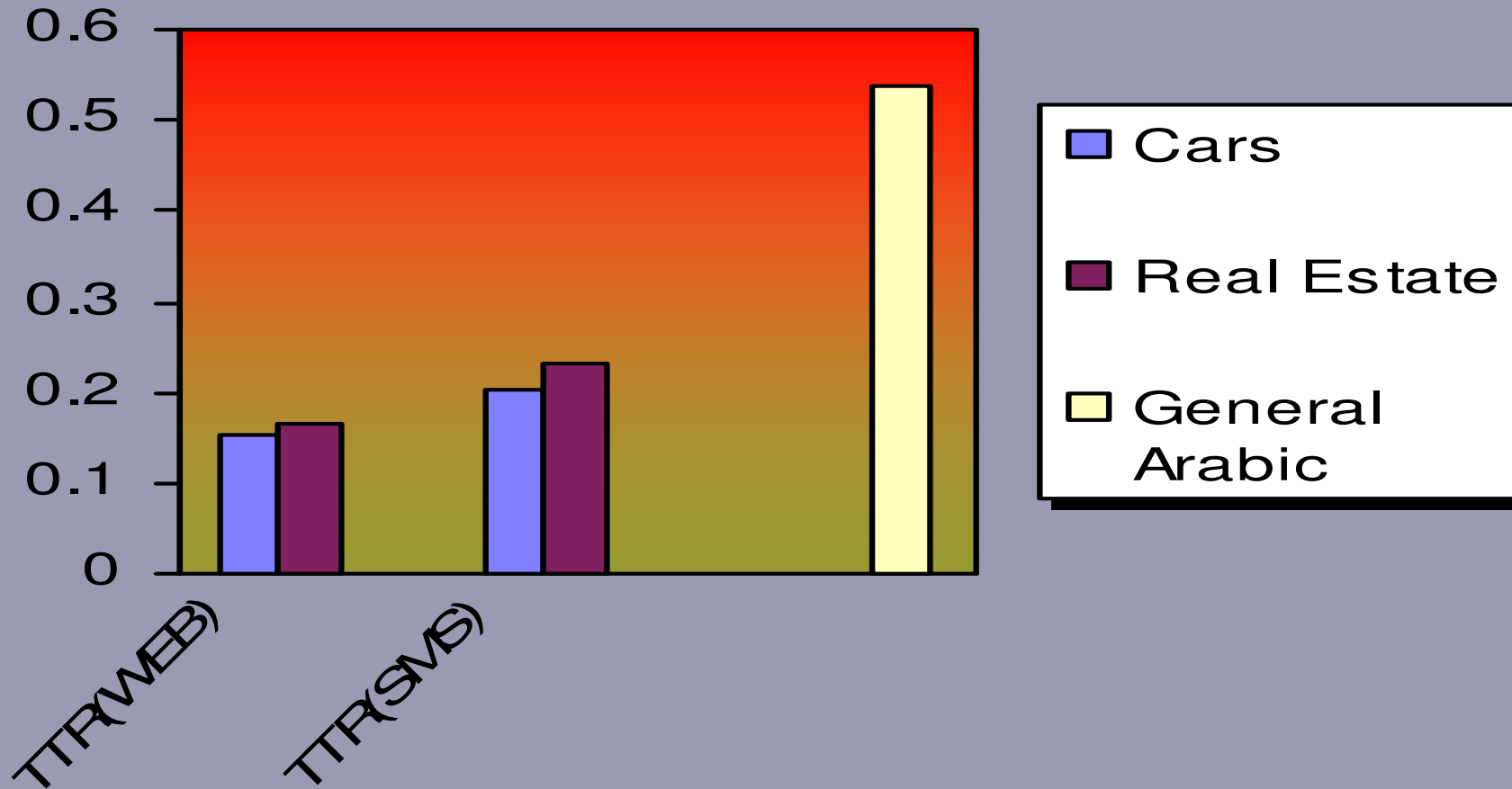
# Sublanguage: examined SMS-based corpus



# Sublanguage: examined SMS-based corpus

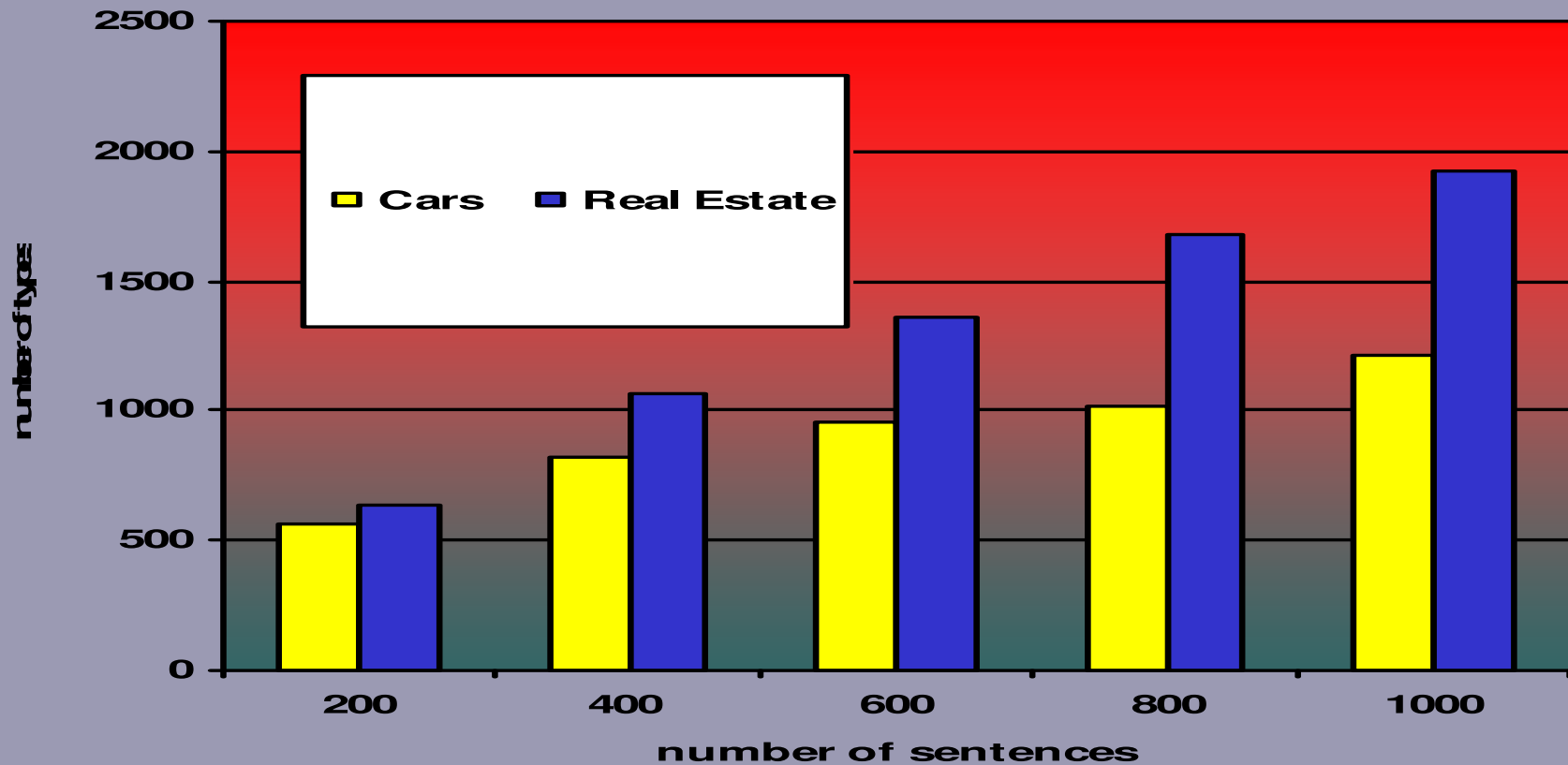


# Sublanguage: examined SMS-based corpus



# Sublanguage: examined SMS-based corpus

number of new types for each 200 sentences



# Sublanguage: Lexical characteristics

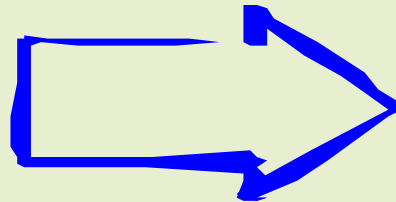
special terminology

للإيجار مطلوب شقة أرضية مع حديقة  
واسعة وكراج ومدخل مستقلين جديدة 2\_3 نوم في  
الرايية أم السماق شارع مكة الشميساني  
المدينة الرياضية لغاية 3500 سنوي

multi-words concepts

Looking for to rent a ground apartment with a large garden and a garage with private entrance new 2-3 bedrooms in Rabiah, Om Alsumaq Mecca street Alshumesani Sports city up to 3500 annual

Named entities



lexical lookup & specialized dictionary

numerical values

techniques for handling spelling variations, segmentations

spelling systems

# Sublanguage: syntactic characteristics

different structures

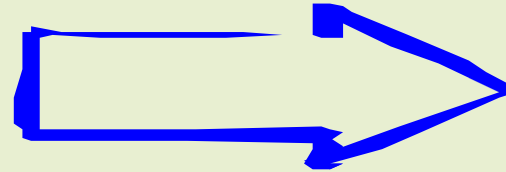
different styles: telegraphic,  
fully formed

Omissions

Syntax-based  
parsing is not  
useful

techniques used for  
semi-structured text  
are not useful

the need for a  
semantic  
description

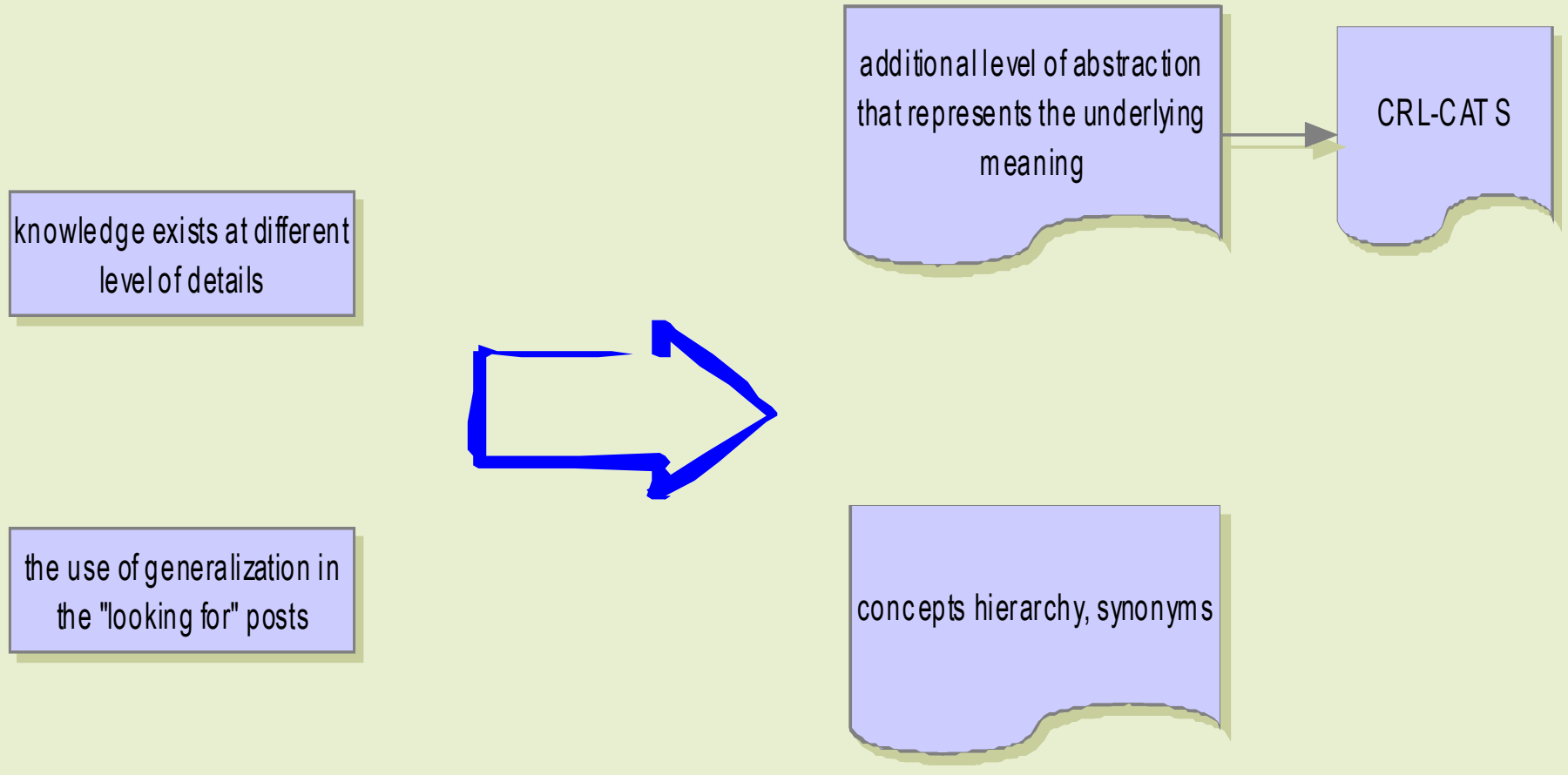


سياره اوبل فكترا للبيع موديل 2003 فل اويشن	Opel Vectra car for sale year 2000 full option
للبيع سيارة بي ام دبليو 520 لون زيتي فحص كامل م 89 فل عدا الفتحة مرخصه بحال ممتازه بسعر 8500	For sale BM W 520 color dark green full check year 89 full except sunro of licensed in a good condition with a price 8500.
أوبل أسترا ستيشن لون أحمر ( بورفتحه سنترزجاج ومريات كهرباء ) فحص للبيع	Opel Astra station color red (power sunro of Center Electrical windows and mirrors check for sale

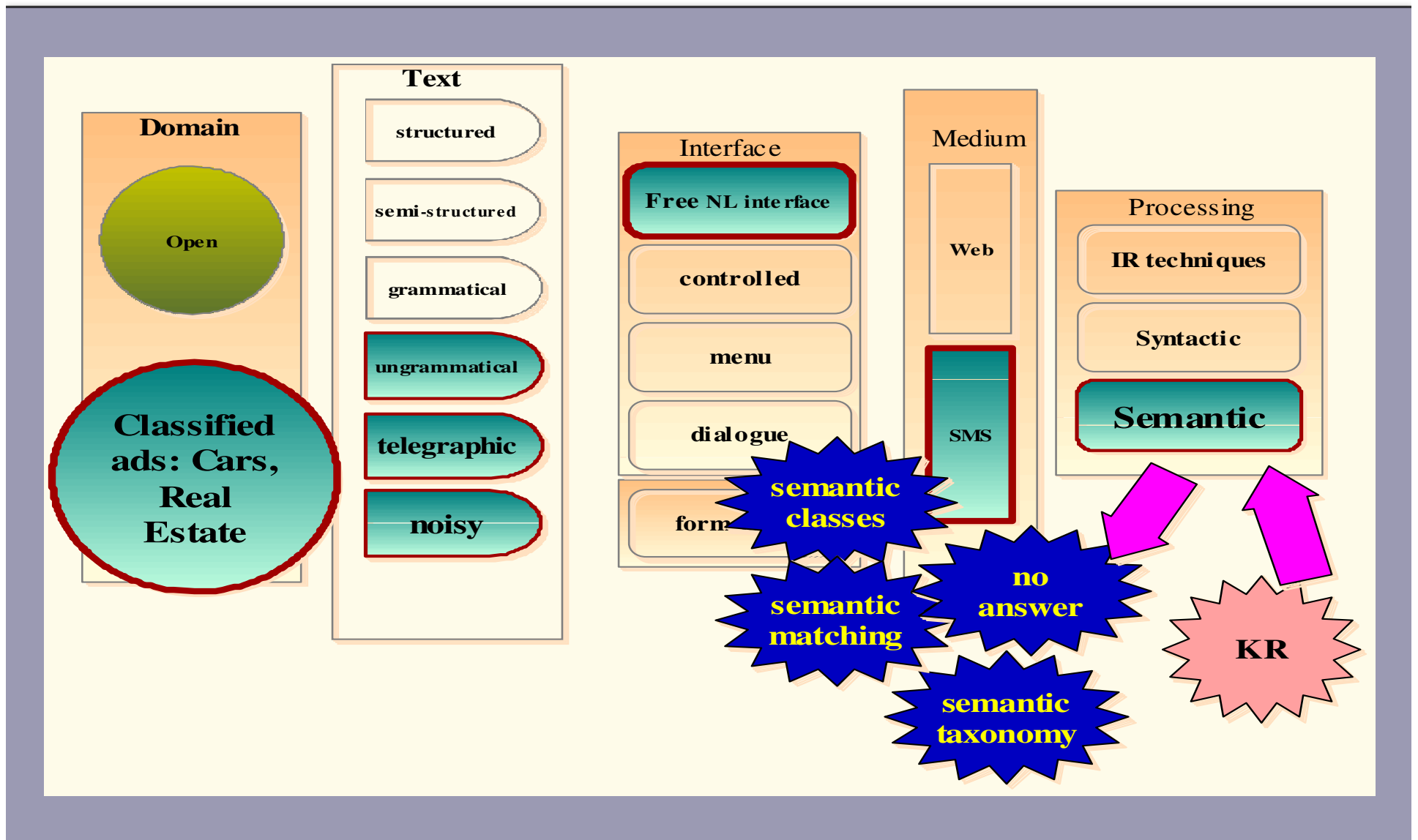
# Sublanguage: semantic characteristics

“looking for a CIVIC” or “A Japanese Honda Civic car for sale”.

“looking for a French car”, “looking for a villa in West Amman” or “looking for economical car”.



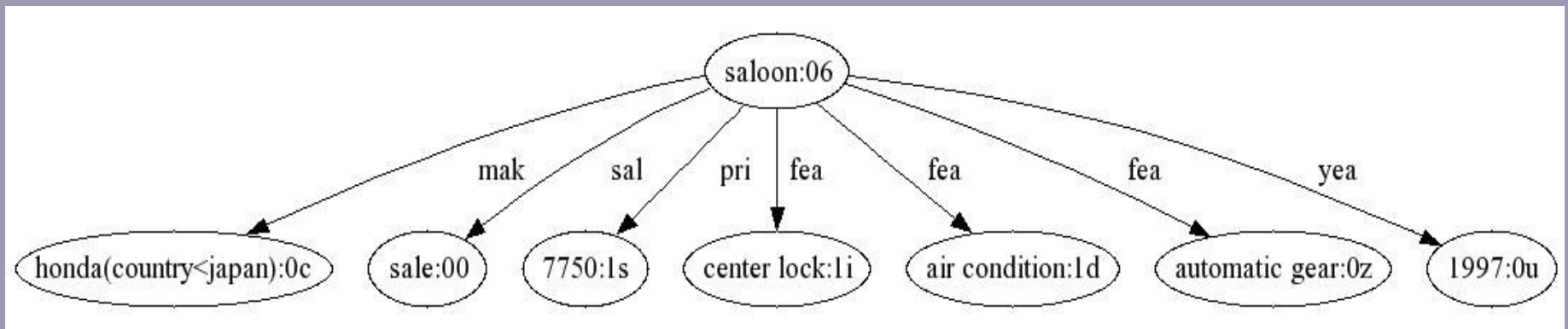
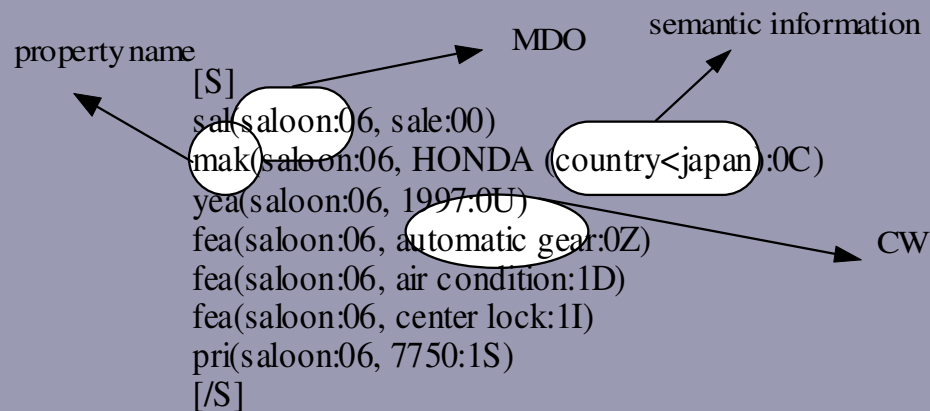
# Sublanguage analysis: outcome



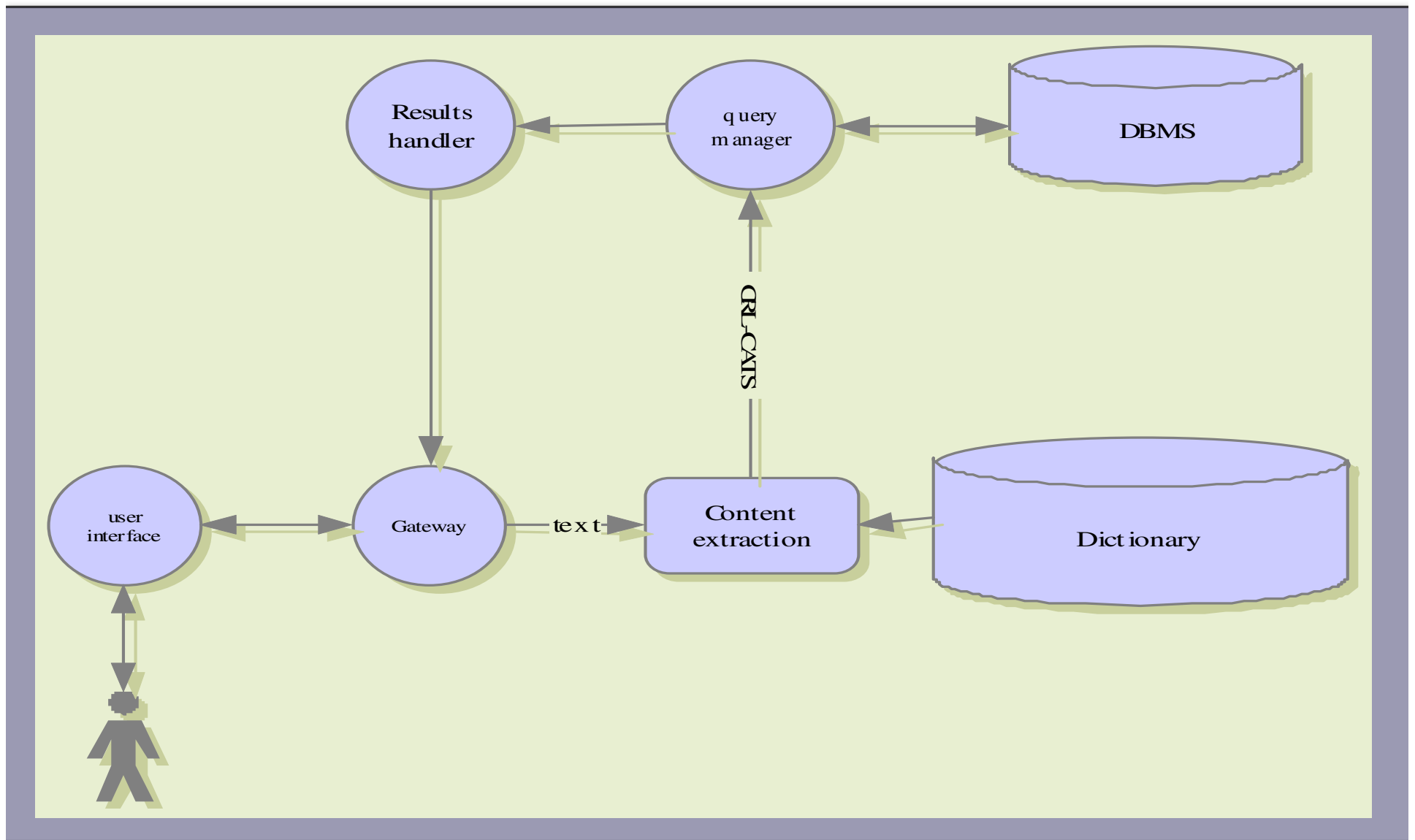
# CE: KR (CRL-CATS)

للبيع سيارة هوندا موديل ١٩٩٧ جير اوتوماتيك مكيف سنتر بسعر  
٧٧٥٠ دينار

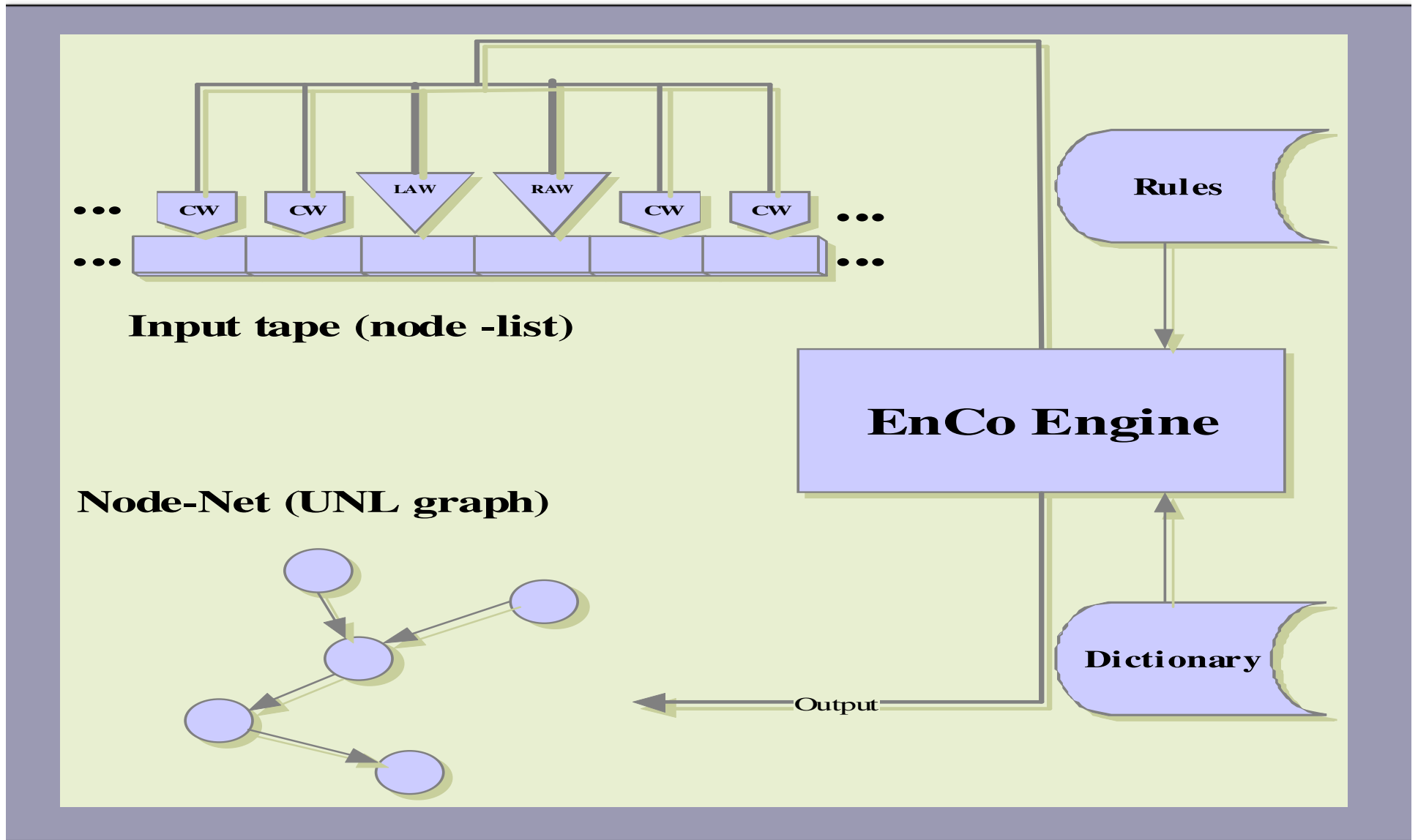
*For sale Honda year 1997 Automatic transmission Air  
condition Center Lock price 7750 Dinar*



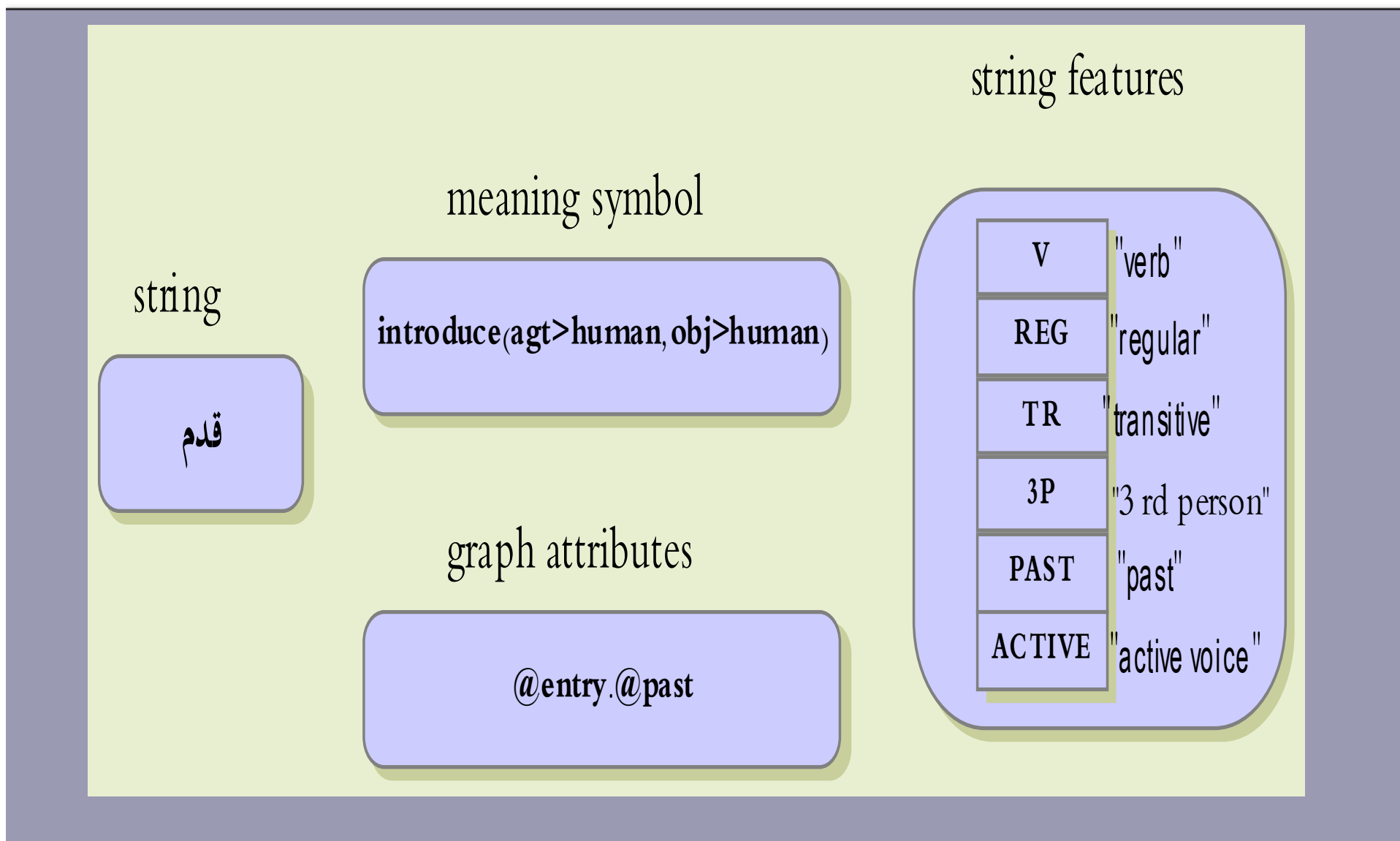
# CE: CATS architecture



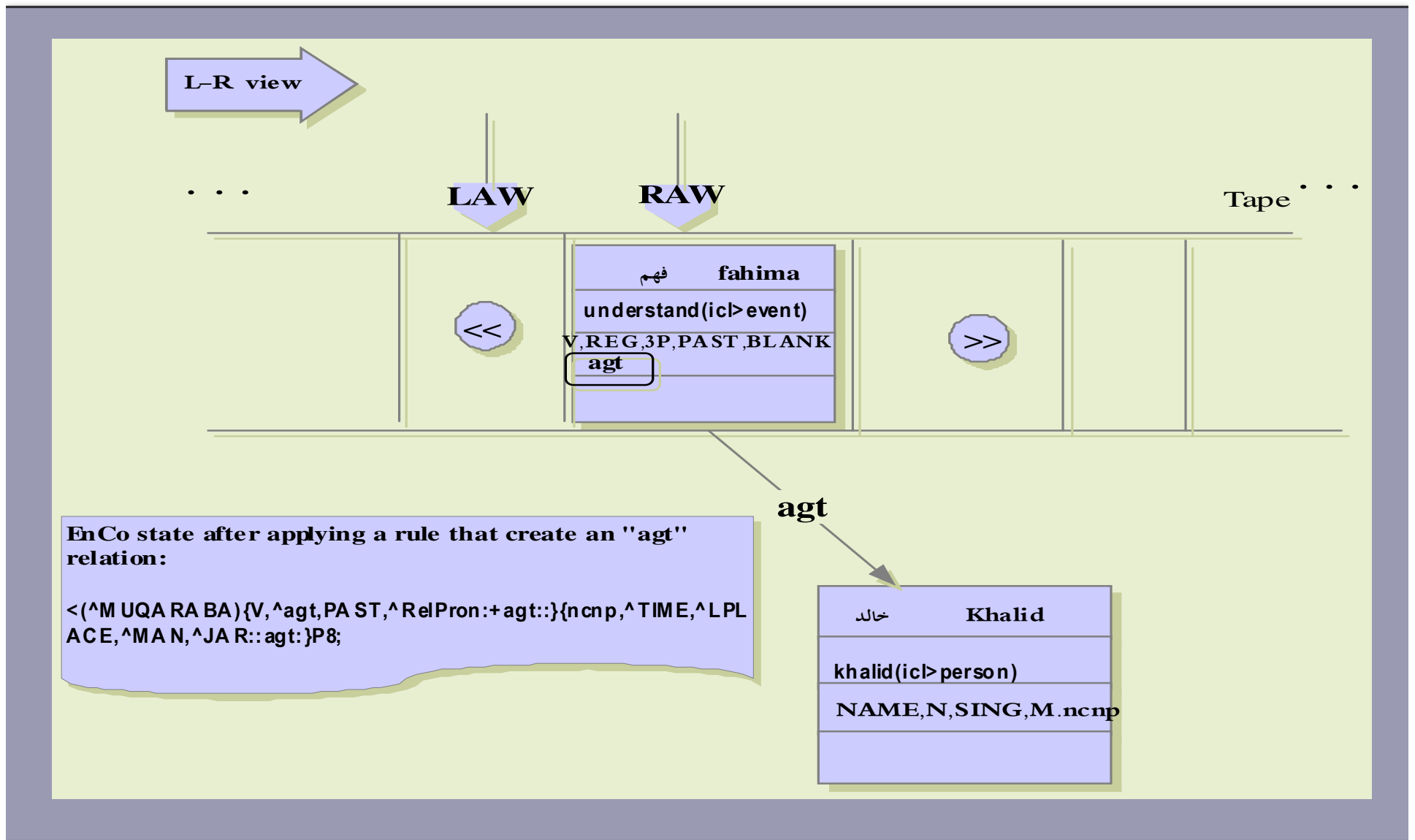
# The structure of EnCo



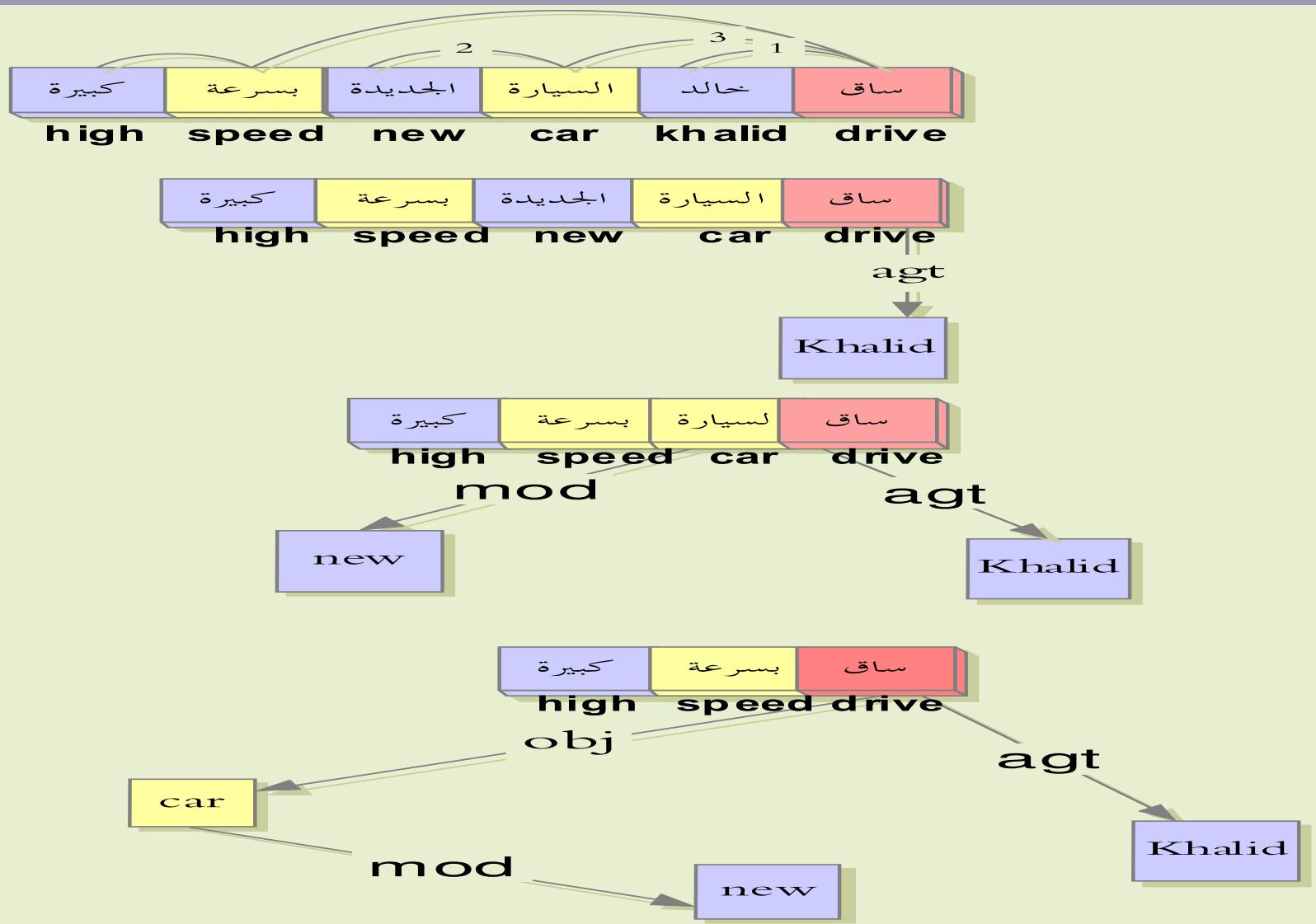
# Example of a node in the tape or the node-net



# EnCo: basic operations

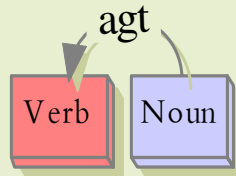


# EnCo: overall analysis strategy

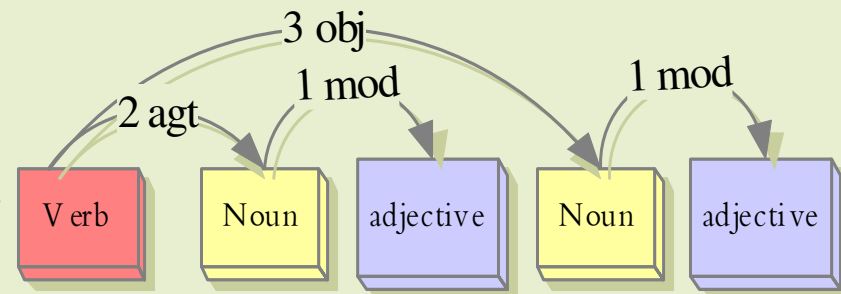
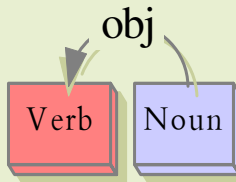


# EnCo rules as a low level formalism for parsing

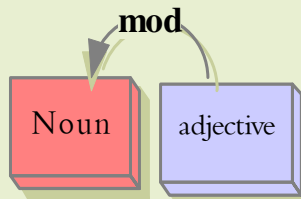
<{V,^agt:agt:}{N::agt:}P8;



<{V,^obj:obj:}{N::obj:}P7;

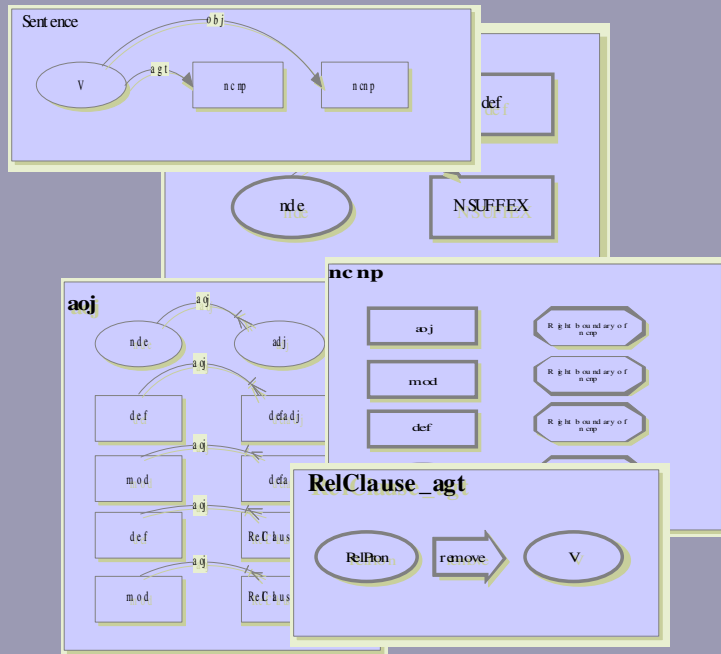


<{N:adj\_added:}{adj::mod:}P11;

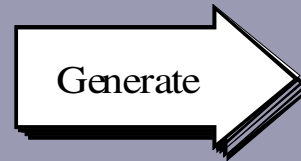


a simplified dependency representation of  
a verbal sentence in Arabic

# The Lingware: Daoud's EnCo Diagrams (DED)



DED diagrams



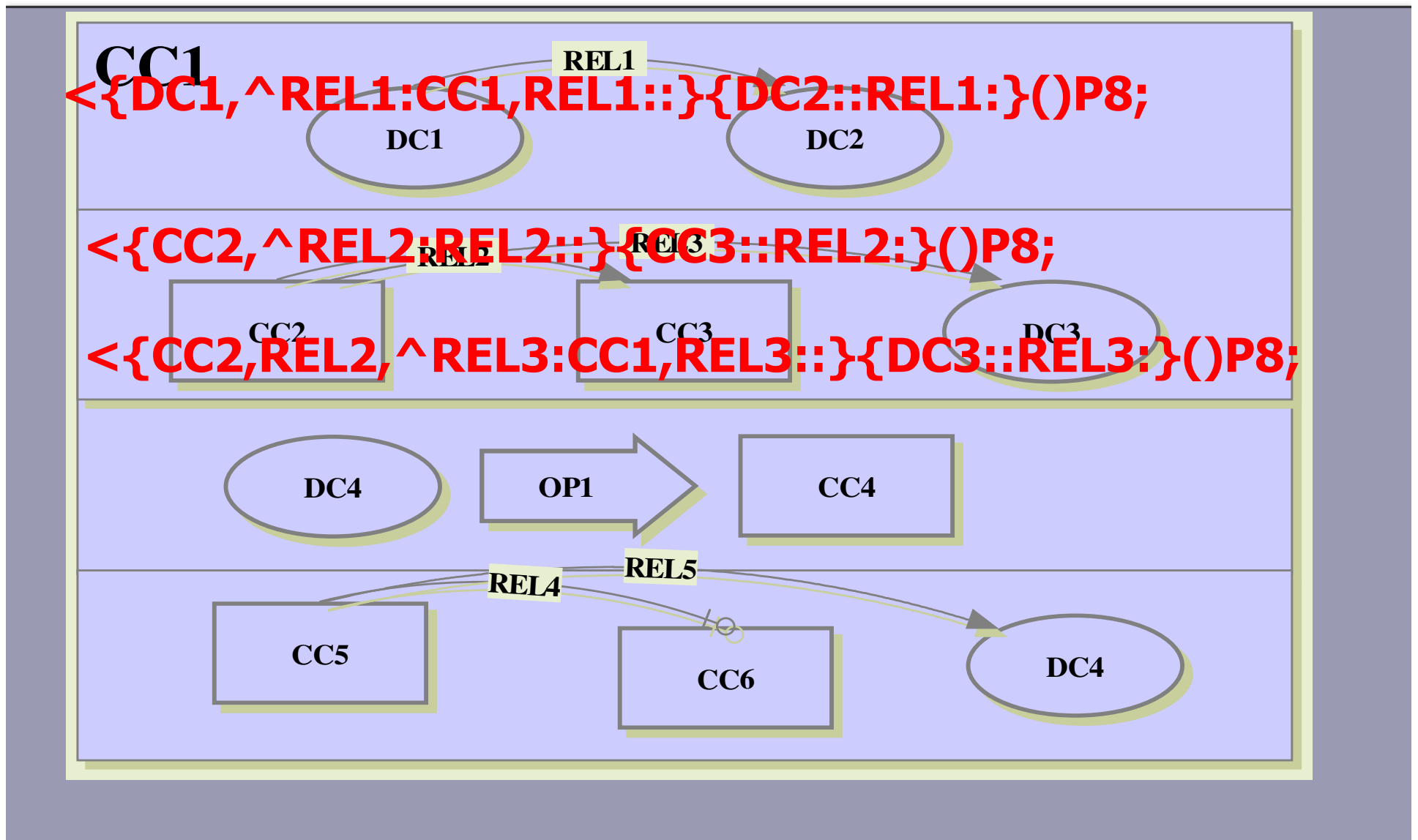
```

=====
+ {::} {BLK:;} ()P255;
=====
R{sentence,^complete:complete,&@entry:} {STAIL:;} ()P5;
=====
R{::} {::} ()P1;
=====
< {V,^agt:agt:} {ncnp:agt:} ()P8;
< {V,^agt:^objobj:sentence:} {ncnp:obj:} ()P7;
=====
L{aoj,^ncnp:ncnp:} {^AL,^V,^RelPron:;} ()P9;
L{mod,^ncnp:ncnp:} {^AL,^V,^RelPron:;} ()P9;
L{nde,^ncnp:ncnp:} {^AL,^V,^RelPron:;} ()P9;
L{def,^ncnp:ncnp:} {^AL,^V,^RelPron:;} ()P9;
=====
< {nde:^nde,aoj:aoj:} {adj:;} ()P10;
< {def:aoj:aoj:} {def:adj:;} ()P10;
< {mod:aoj:aoj:} {def:adj:;} ()P10;
< {def:aoj:aoj:} {RelClause_obj:;} ()P10;
< {mod:aoj:aoj:} {RelClause_obj:;} ()P10;
=====
< {nde,^mod:mod:} {def:mod:} ()P11;
< {nde,^mod:mod:} {NSUFFEX,MODREL:mod:} ()P11;
=====
- {FW,AL:;} {nde,^def:-nde,def:;} ()P12;
L{::} {NAME,^def:-NAME,def:;} ()P12;
=====
- {FW,AL:;} {adj,^def:adj:-adj,def:adj:;} ()P11;
=====
< {V,RelClause_agt,^RelClause_obj:RelClause_obj:} {ncnp:obj:} ()P11;
=====
- {RelPron:RelClause_agt:} {V:;} ()P12;
=====

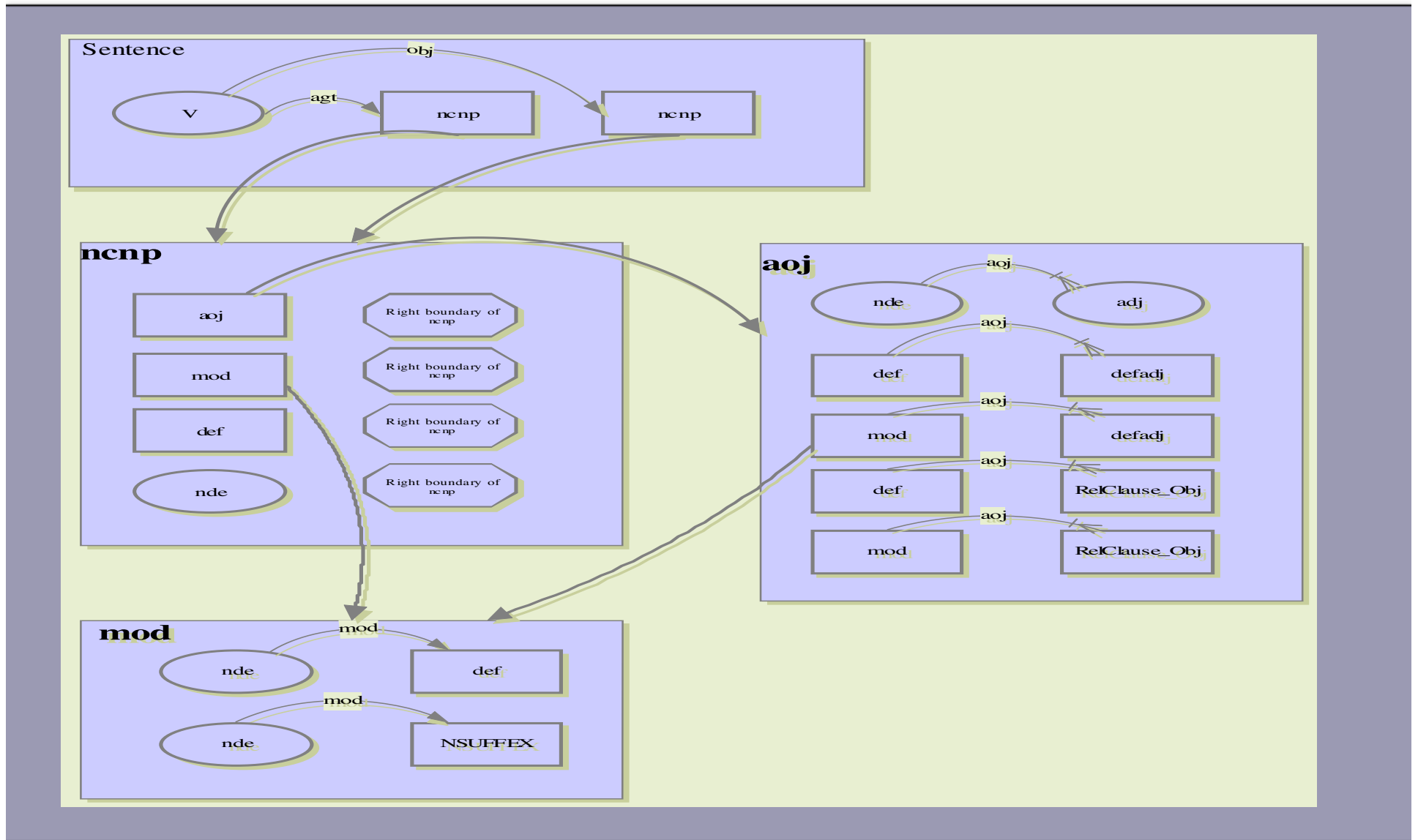
```

EnCo rules

# The lingware: mapping procedures



# The lingware: an example



# The lingware: testing the DED

فهم خالد الذكي الذي قابل محمد القوي الذي قرأ كتاب داود المفيد الرائع الدرس

The clever Khalid who met the strong Moh'd who read Daoud's fantastic useful book understood the lesson.

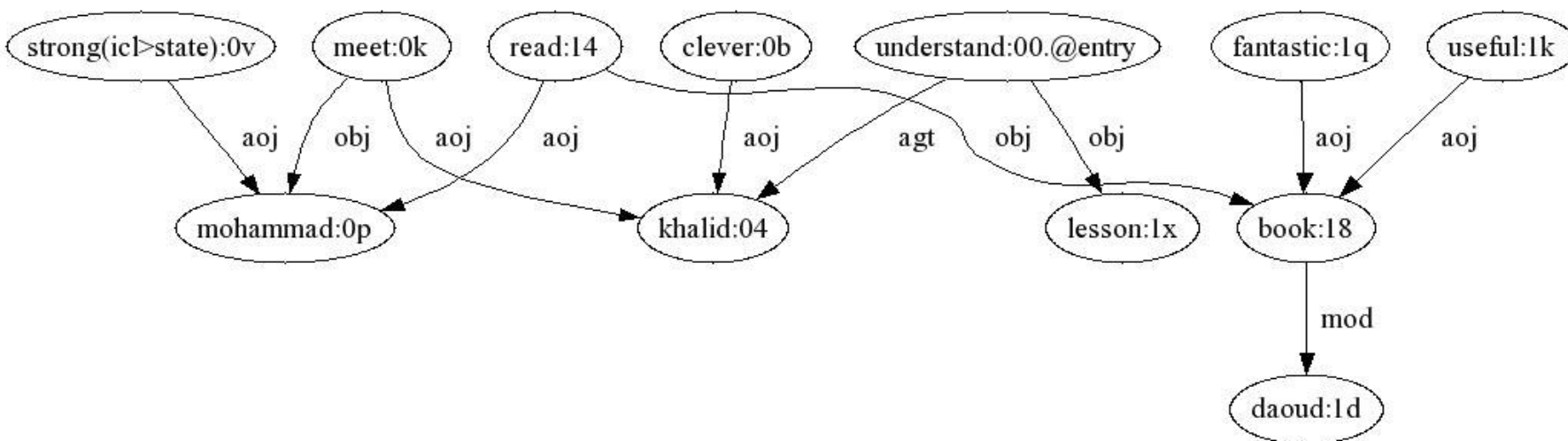
[S]

aoj(clever:0B, Khalid:04)

aoj(strong(icl>state):0V, mohammad:0P)

mod(book:18, daoud:1D)

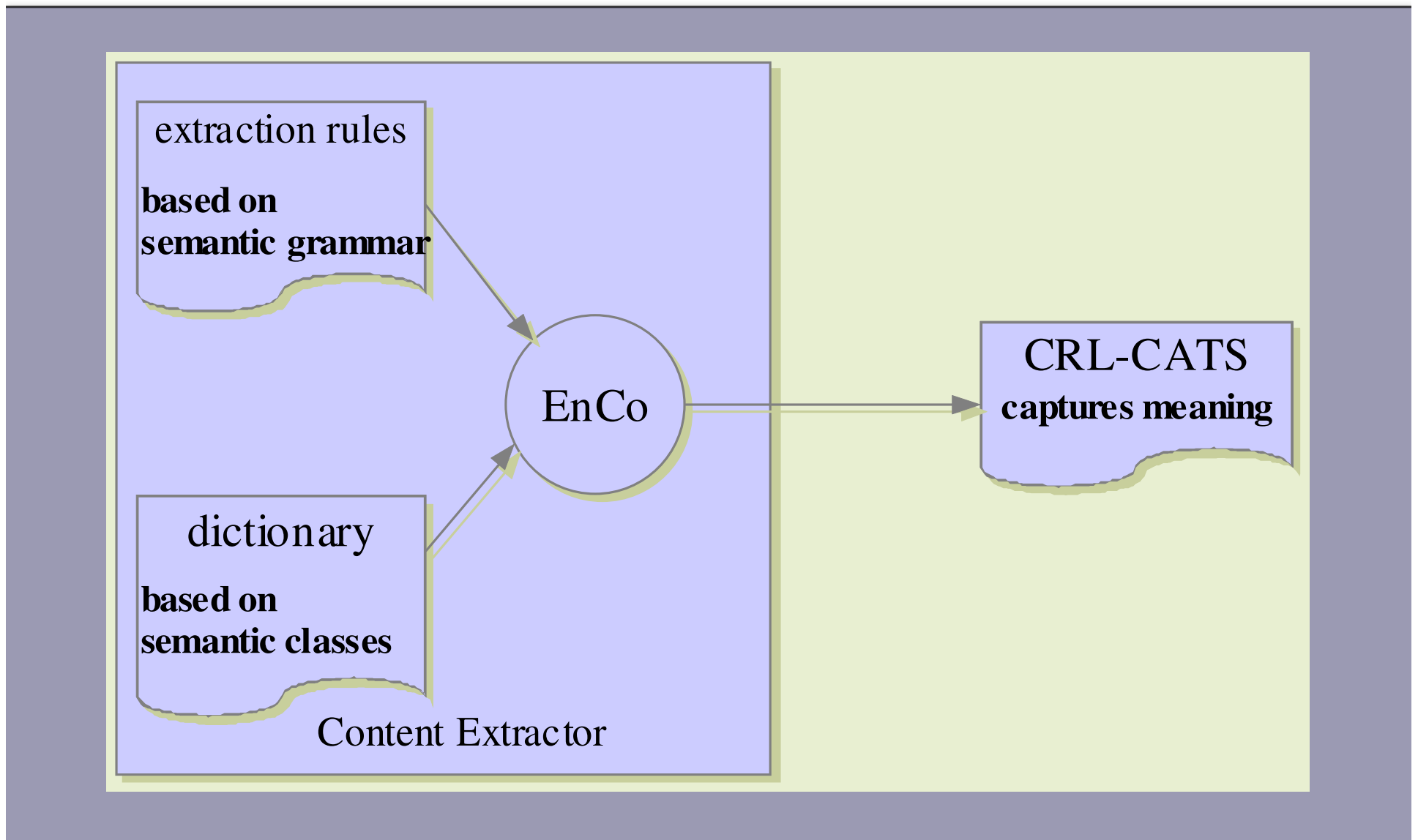
aoj(useful:1K, book:18)



# Outline

- Short proof of “necessity” part
- **Proof of “possibility” part with some original contributions**
  - Sublanguage Analysis
  - **Content Extraction**
    - CE in EnCo
    - DED methodology
    - **CE: the case of CATS**
  - The QA component
  - Evaluation and results
- Conclusions and perspectives

# Content extraction: architecture



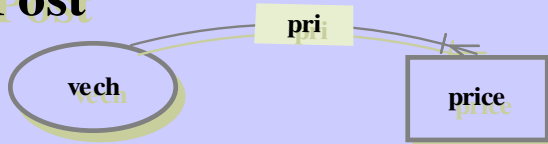
# Content Extraction: the use of semantic grammar

## The Price Knowledge component

السعر 500 دينار	The price is 500 dinar	Pri(flat,500@dinar)
500 دينار	500 dinar	Pri(flat,500@dinar)
السعر 500	The price is 500	Pri(flat,500)
لا يزيد السعر عن 500 دينار	Not more the price 500 dinar	Pri(flat,500@dinar@less)
السعر لا يزيد عن 500	The price is not more 500	Pri(flat,500@less)
500 دينار و اكثر	500 dinar and above	Pri(flat,500@dinar@more)
500 و مافوق دينار	500 and above dinar	Pri(flat,500@dinar@more)

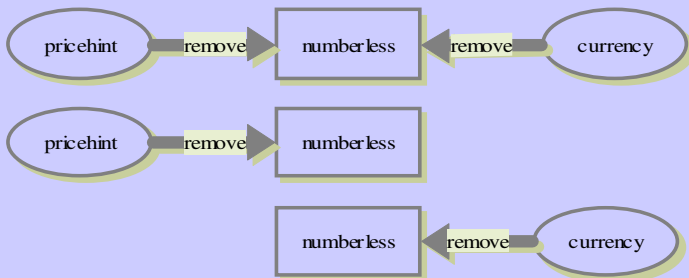
# Content Extraction: the use of DED

## CarPost

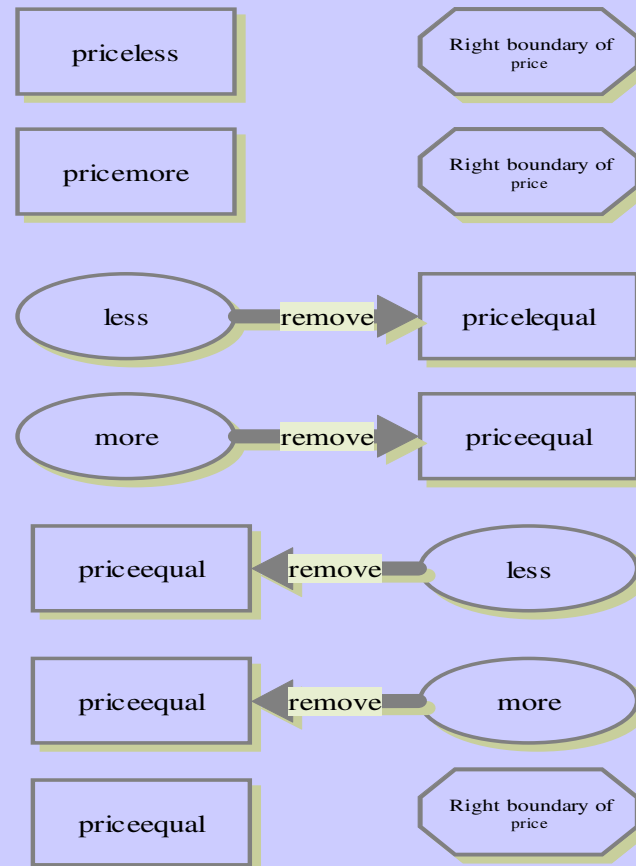


<{vech:CarPost,pri:}{price::pri;}P10;

## priceless



## price



# Content Extraction: the dictionary

- Compensates for lexical inconsistency by connecting words to concepts (CWs).
- Provides the semantic information needed for reasoning.
- The number of CWs is 10828.
- The number of lexical forms is 30982.

```
[دبلیو بی ام]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;  
[دبلیو بی ام]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;  
[بی.ام.دبلیو]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;  
[بی.ام.دبلیو]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;  
[دبلیو بی ام]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;  
[دبلیو بی ام]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;  
[بی.ام.دبلیو]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;  
[بی.ام.دبلیو]{}"B.M.W(country<germany,country<europe)"(make,nmodel,car) <A,3,3>;
```

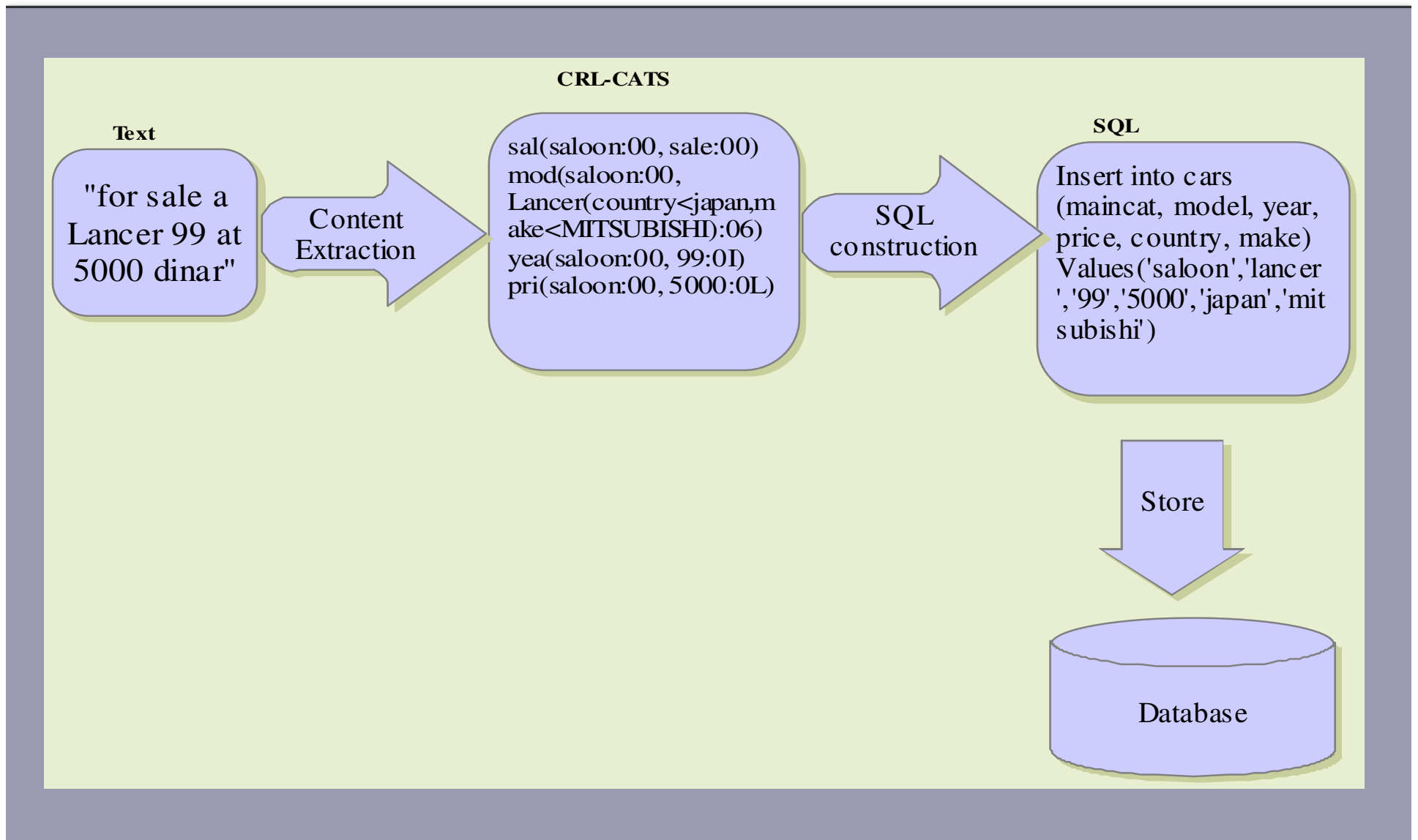
# Extraction rules

- **Based on semantic description.**
- **Numerical values recognition**
- **Named entities extraction**
- **Extraction of relation**
- **Extraction of significant information only**
- **710 rules for both Cars and Real Estate domain.**

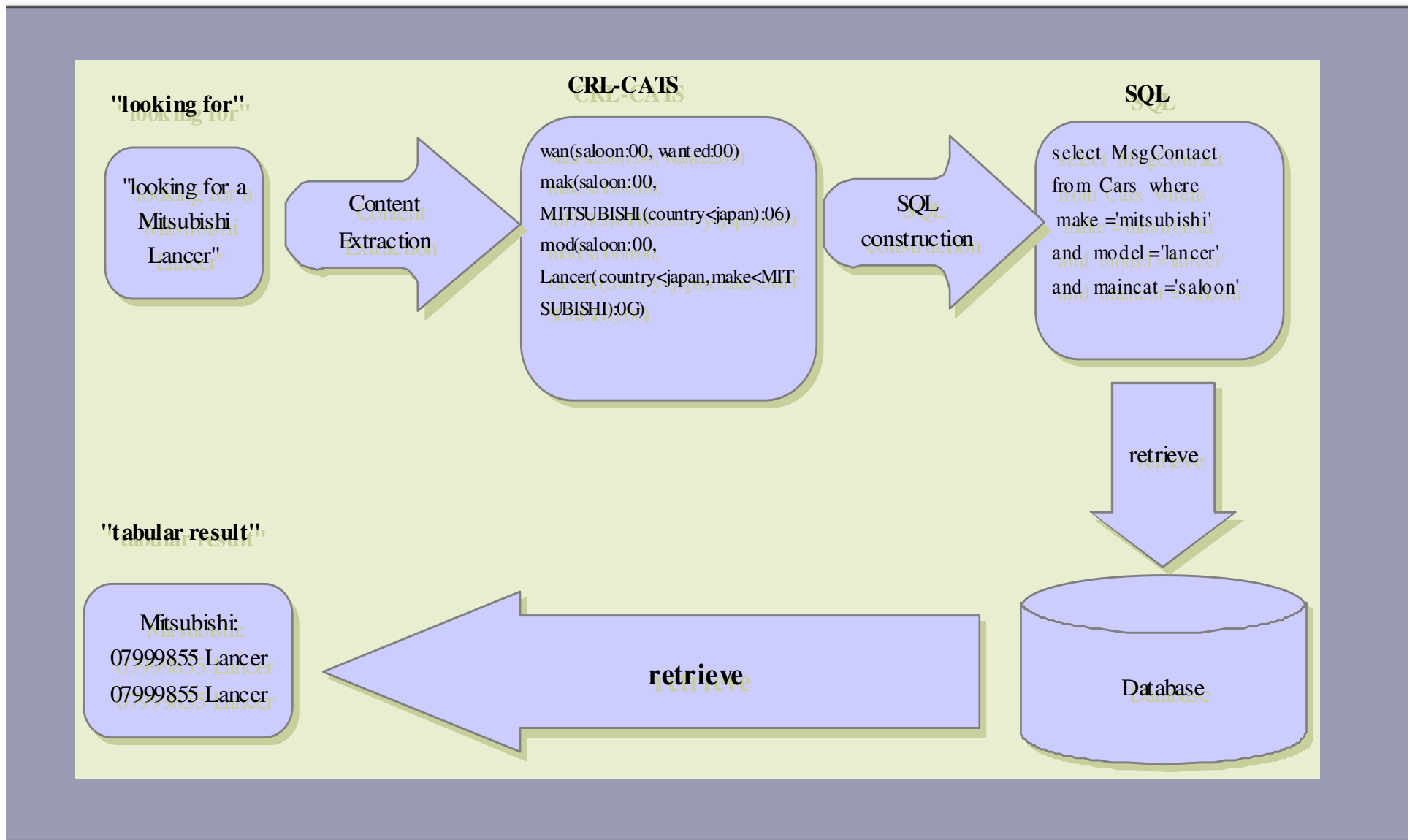
# Semantic-based matching

- **CE handles different causes of mismatch at the local level.**
- **Other sources of mismatch are handled at the global level:**
  - **Granularity**
  - **Generalization**
  - **No answer situations**
- **Exact matching is not sufficient. Semantic matching is used.**
- **DB has to be designed with an inference mechanism.**

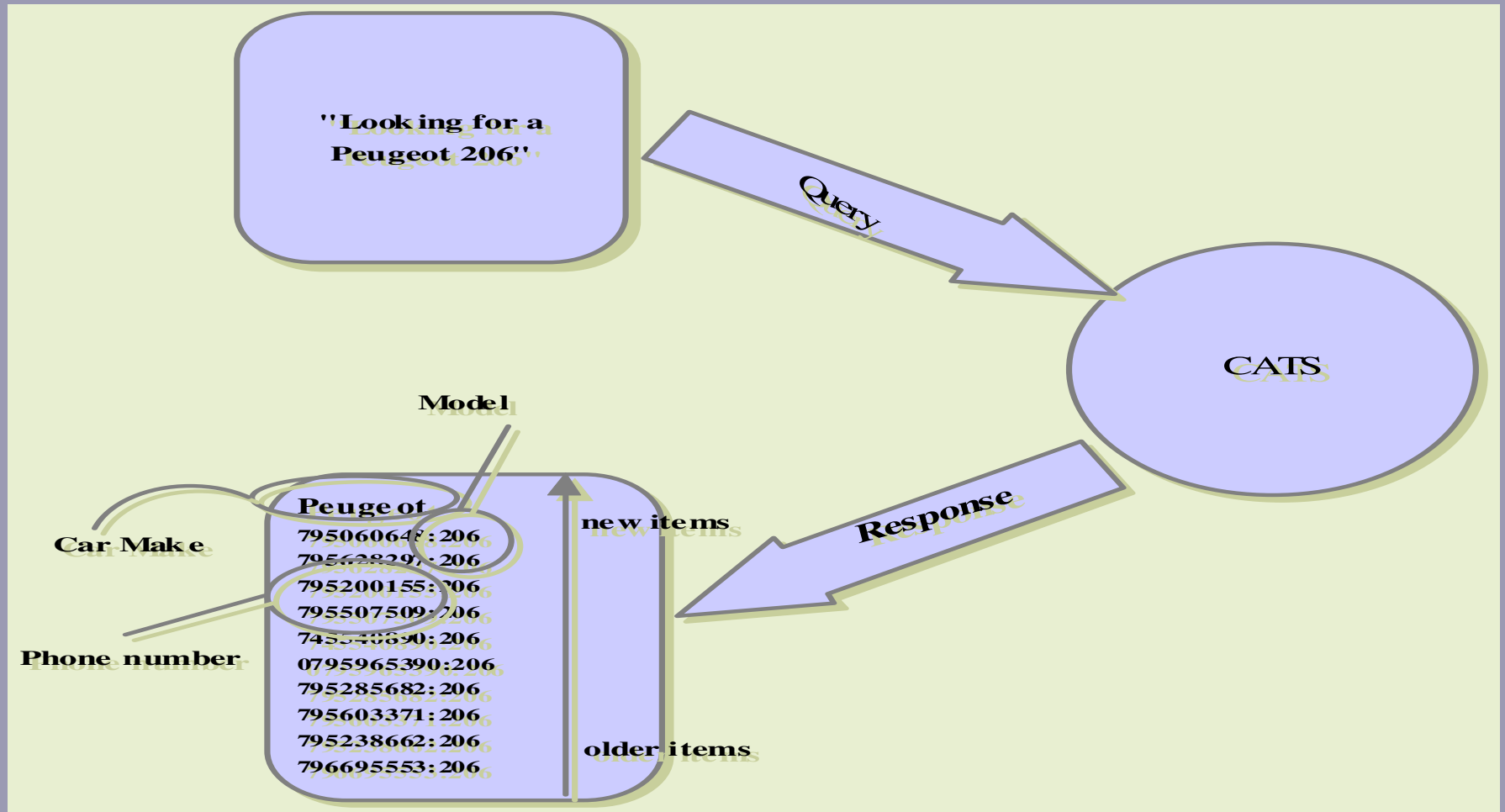
# Information flow of "sell" post



# Information flow of "looking for" post



# QA: Generating Responses



# QA: Database

**Horizontal table representation**

Object-id <sub>i</sub>	attr1 <sub>i</sub>	attr2 <sub>i</sub>	attr3 <sub>i</sub>
1 <sub>i</sub>	a <sub>i</sub>	b <sub>i</sub>	null <sub>i</sub>
2 <sub>i</sub>	null <sub>i</sub>	c <sub>i</sub>	d <sub>i</sub>
3 <sub>i</sub>	null <sub>i</sub>	null <sub>i</sub>	a <sub>i</sub>
4 <sub>i</sub>	b <sub>i</sub>	null <sub>i</sub>	d <sub>i</sub>

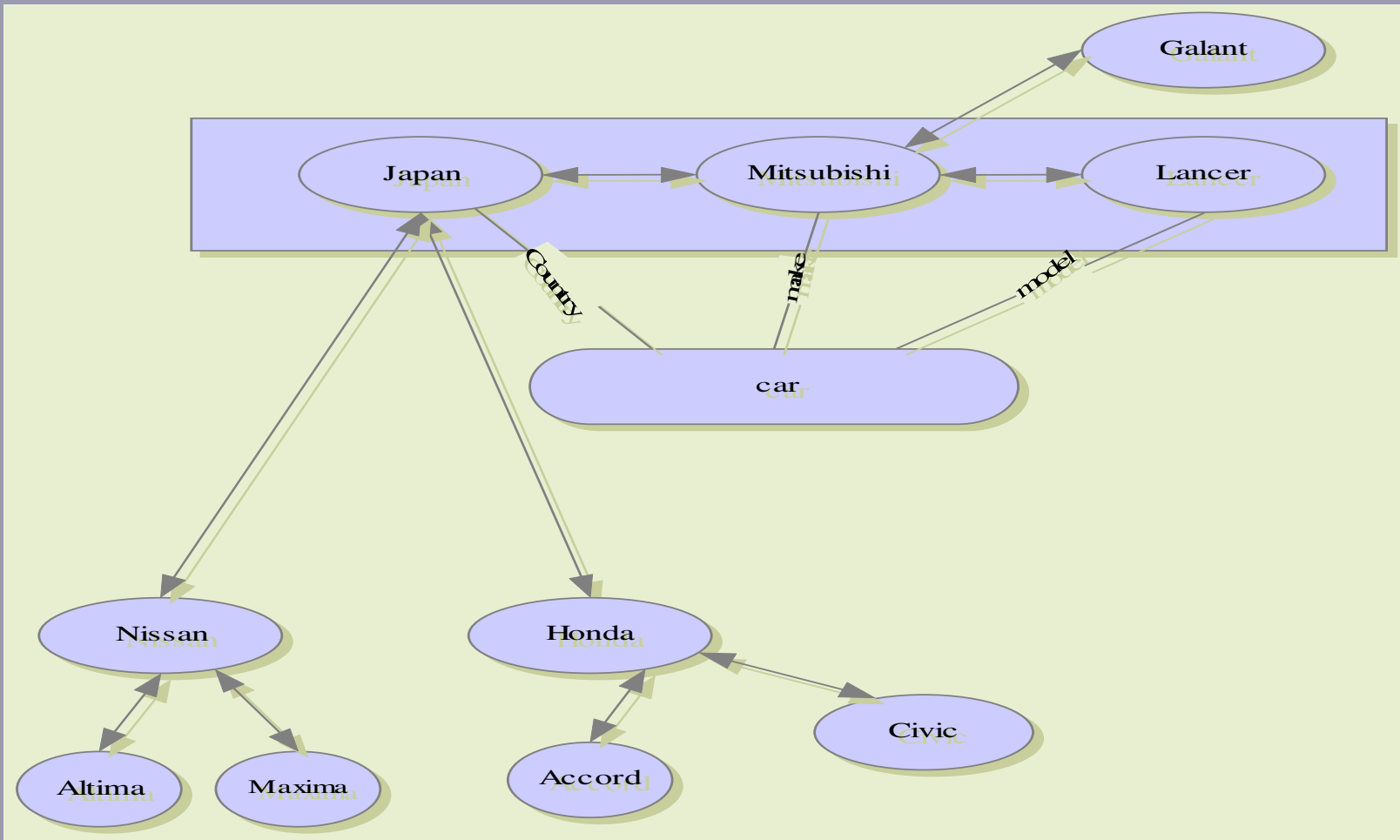
**Vertical table representation**

Object-id <sub>i</sub>	key <sub>i</sub>	value <sub>i</sub>
1 <sub>i</sub>	attr1 <sub>i</sub>	a <sub>i</sub>
1 <sub>i</sub>	attr2 <sub>i</sub>	b <sub>i</sub>
2 <sub>i</sub>	attr2 <sub>i</sub>	c <sub>i</sub>
2 <sub>i</sub>	attr3 <sub>i</sub>	d <sub>i</sub>
3 <sub>i</sub>	attr3 <sub>i</sub>	a <sub>i</sub>
4 <sub>i</sub>	attr1 <sub>i</sub>	b <sub>i</sub>
4 <sub>i</sub>	attr3 <sub>i</sub>	d <sub>i</sub>

# QA: Database

<b>column</b>	<b>meaning</b>	<b>source of information</b>
<b>id</b>	post id	system
<b>msgCaller</b>	mobile number of the poster	SMS message
<b>msgDate</b>	data of the message	SMS message
<b>msgTxt</b>	content of the post	SMS message
<b>maincat</b>	MDO	CRL-CATS
<b>make</b>	car maker	CRL-CATS
<b>model</b>	car model	CRL-CATS
<b>country</b>	manufacturer country	CRL-CATS
<b>motor</b>	motor size	CRL-CATS
<b>makeyear</b>	year	CRL-CATS
<b>price</b>	price	CRL-CATS
<b>color</b>	color	CRL-CATS
<b>feature</b>	feature	CRL-CATS
<b>strGUID</b>	Globally Unique Identifier	system
<b>IsBuyer</b>	"sell" or "looking for"	CRL-CATS
<b>IsInactive</b>	active or inactive post	SMS message
<b>SendFlag</b>	indicates if the query is answered or not	system

# QA: Reasoning and semantic-based matching



# QA: Reasoning and semantic-based matching (scenario 1)

**Cars table**

id	msgcaller	maincat	make	model	----	....
1	079667999	saloon	Null	Clio		
2	07989999	saloon	Renault	Clio		
3	07988856	saloon	Null	Megan		
4	079777	saloon	Peugeot	Null		
.....	.....	.....	.....	.....		

**SELECT msgcaller**

**FROM cars**

**WHERE make in**

**(select make from countrymake where country = "France")**

**Or**

**model in (select model from makemodel where make in (select make from countrymake where country = "France" ) );**

CountryMake table:

Country	Make
France	Renault
France	Peugeot
Japan	Honda
...	

MakeModel table:

Make	Model
Renault	Clio
Peugeot	Megan
Honda	Civic

# QA: Reasoning and semantic-based matching (scenario 2)

## Cars table

id	msgcaller	maincat	make	model	Country	MsgTxT
1	079667999	saloon	Renault	Clio	France	for sale a Clio
2	07989999	saloon	Renault	Clio	France	For sale a Renault Clio
3	07988856	saloon	Renault	Megan	France	For sale a Megan
4	079777	saloon	Peugeot	Null	France	for sale a Peugeot
5	078666	Saloon	Honda	Civic	Japan	for sale a Honda Civic
.....						

## QA: Storing "sell" posts

*"for sale a Lancer 99 at 5000 dinar"*

[S]

sal(saloon:00, sale:00)

mod(saloon:00, Lancer(country<japan,make<MITSUBISHI):06)

yea(saloon:00, 99:0I)

pri(saloon:00, 5000:0L)

[/S]

**Insert into cars (maincat, model, year, price, country, make)**

**Values('saloon','lancer','99','5000','japan','mitsubishi')**

## QA: Processing of “looking for” posts

*“looking for a Mitsubishi Lancer”*

[s]

wan(saloon:00, wanted:00)

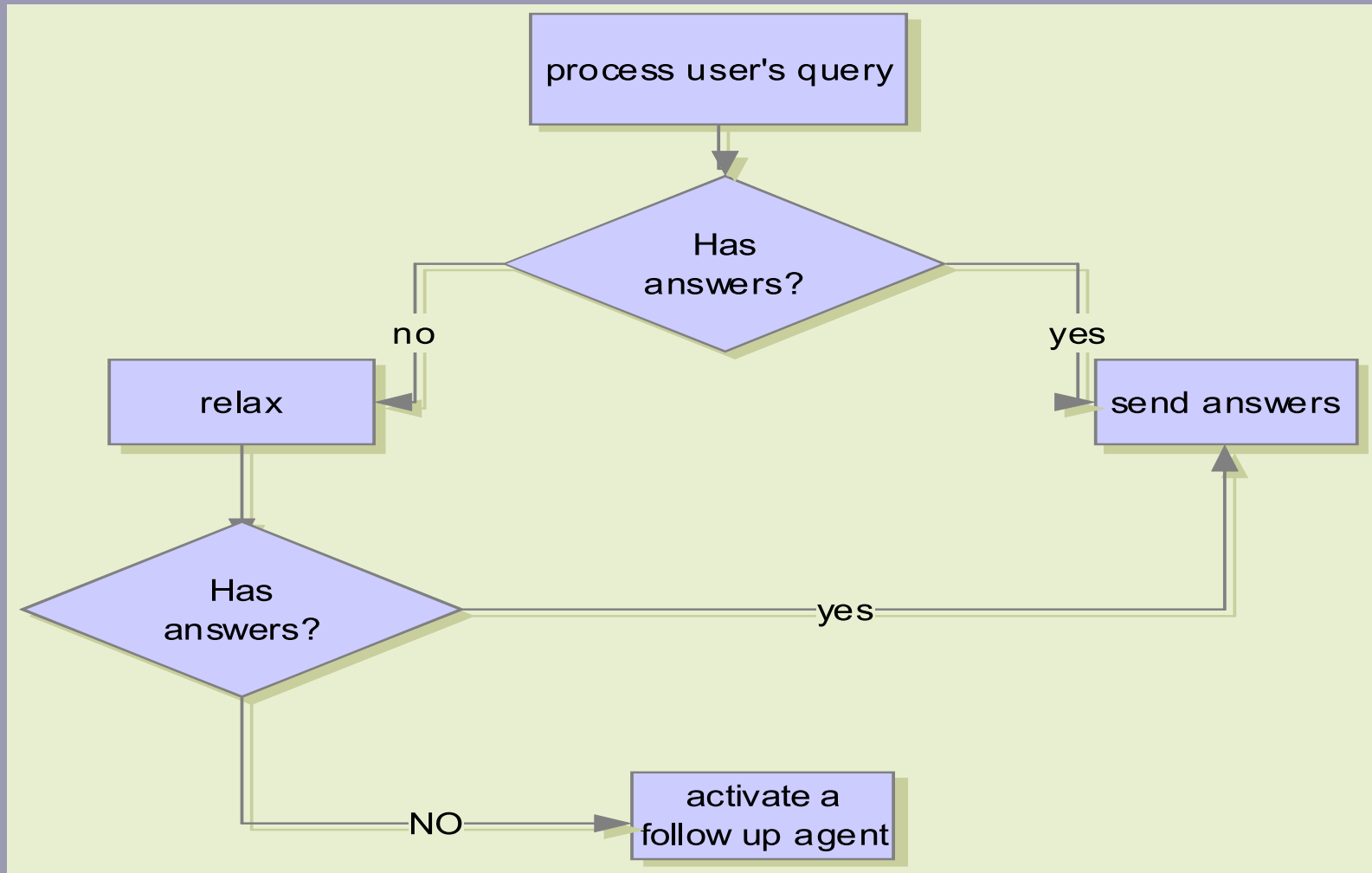
mak(saloon:00, MITSUBISHI(country<japan):06)

mod(saloon:00, Lancer(country<japan,make<MITSUBISHI):0G)

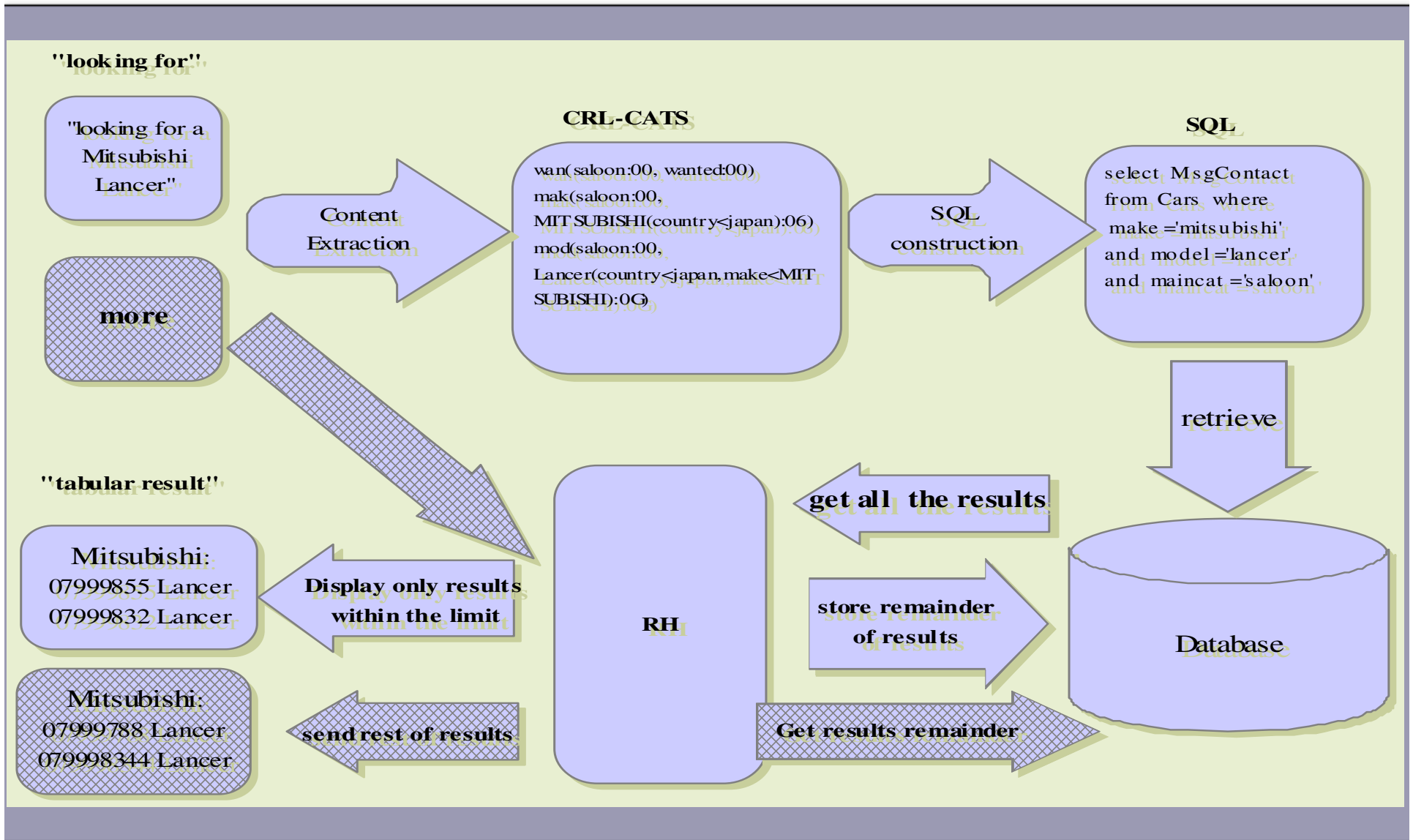
[/S]

```
select MsgCaller from Cars where  
(make = 'mitsubishi' or model = 'lancer')  
and maincat = 'saloon'
```

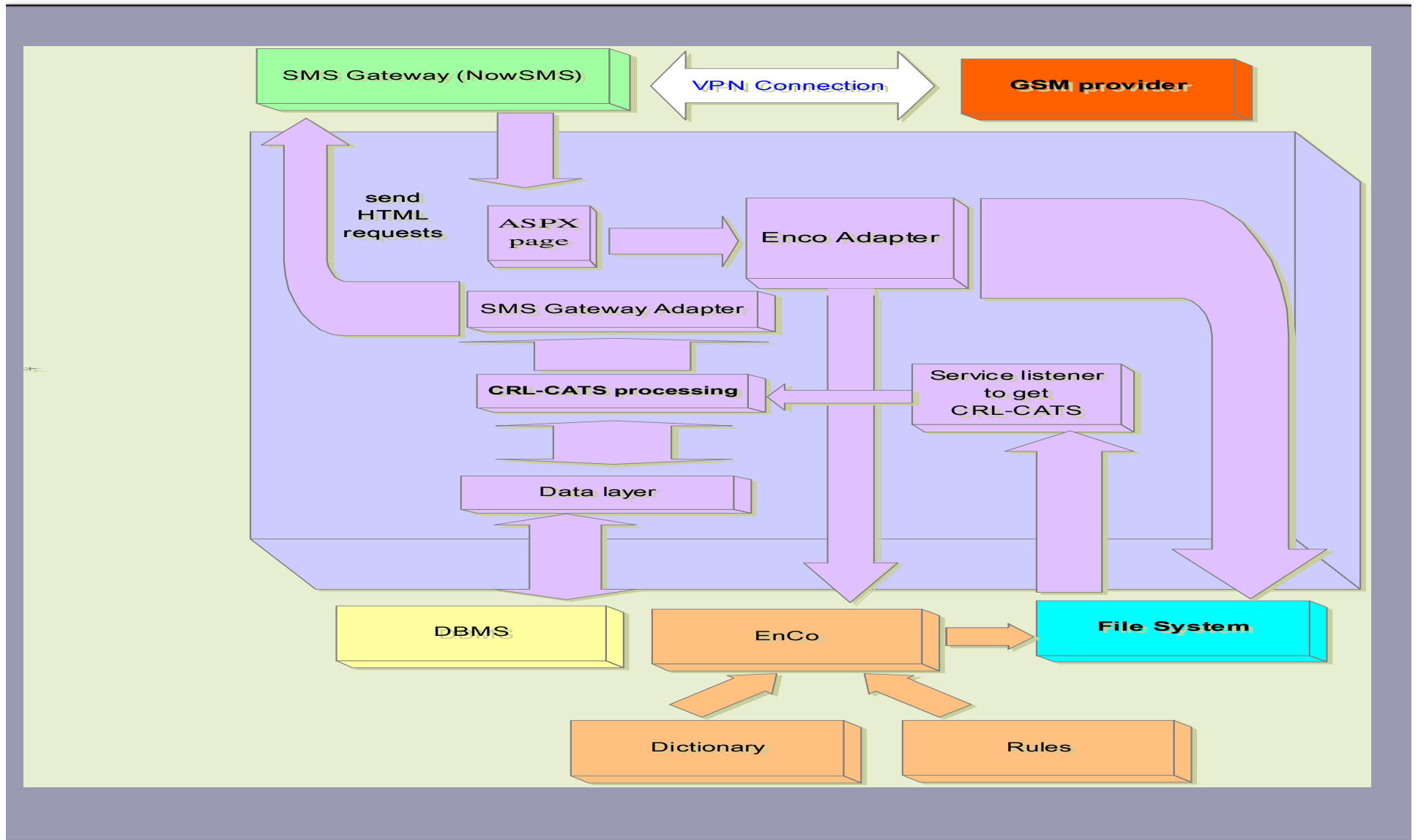
# No answer situations



# Follow up



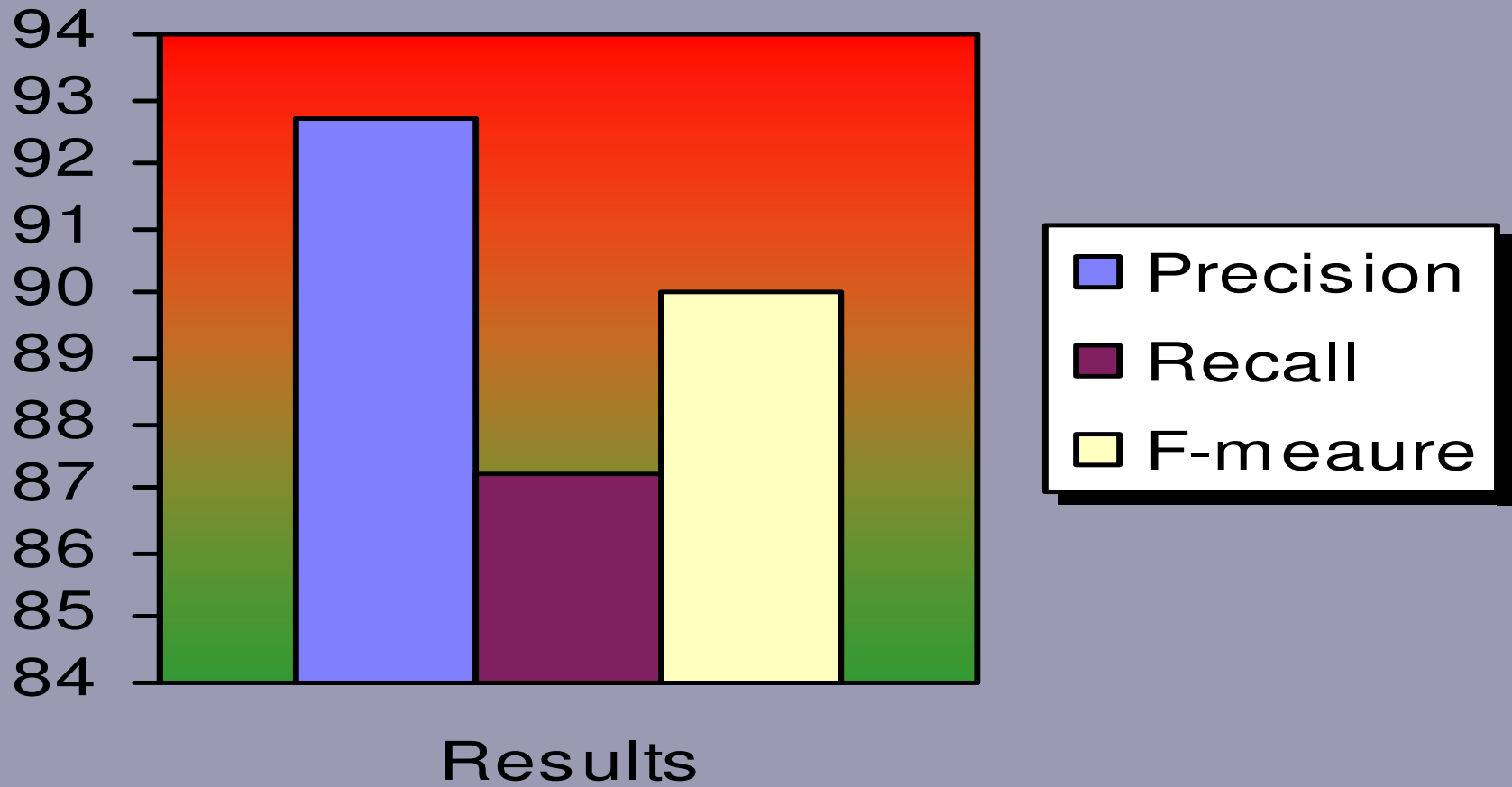
# Operational status : CATS "production" architecture



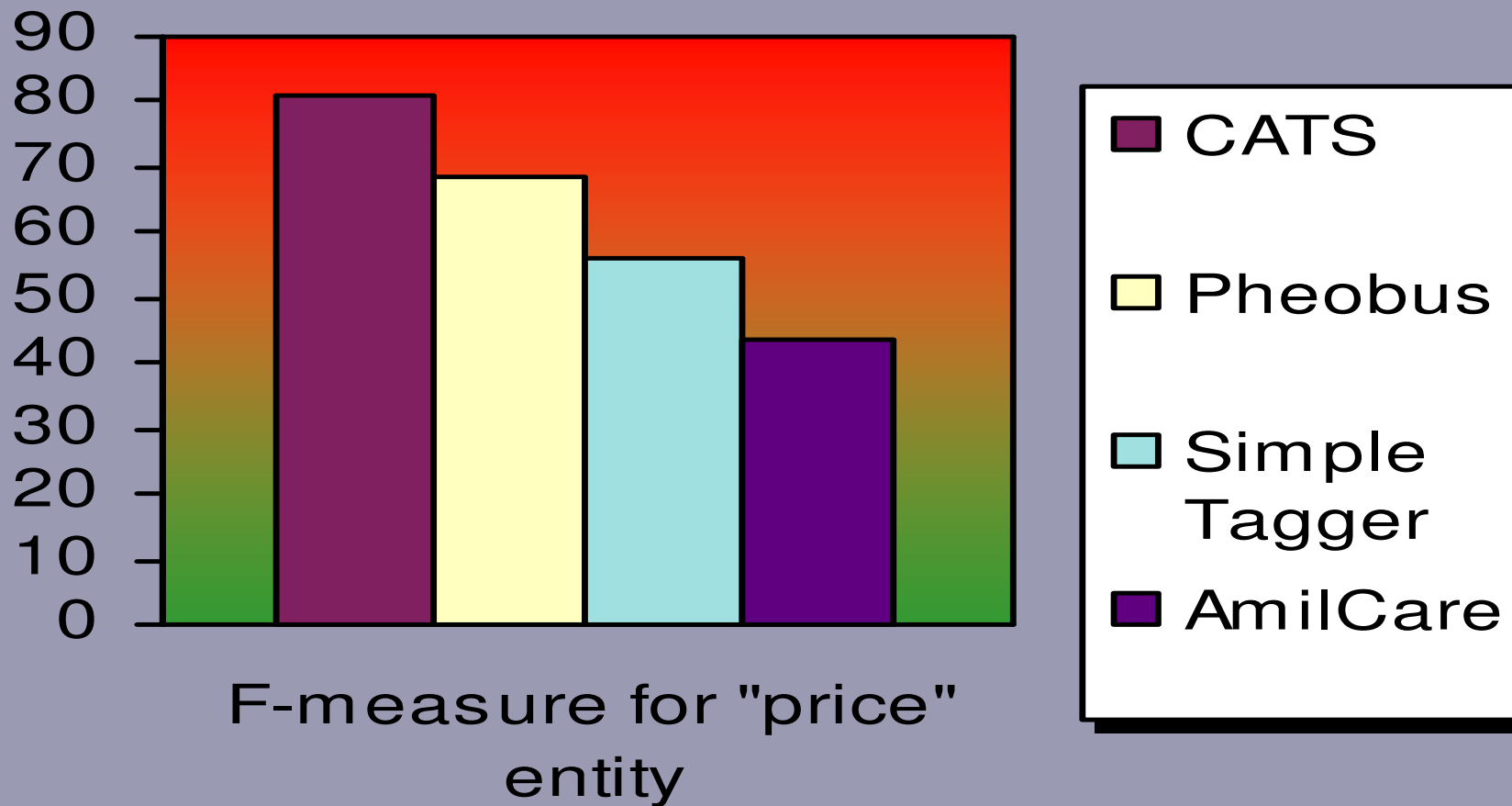
# Estimation of resources used

Task	size	resource (man-hour)
NLP	710 rules+ 31000 entry	2000
Processing system	4000 lines of code	1000
<b>Modular testing</b>		200
<b>Integration</b>		100
<b>Integrating testing</b>		100

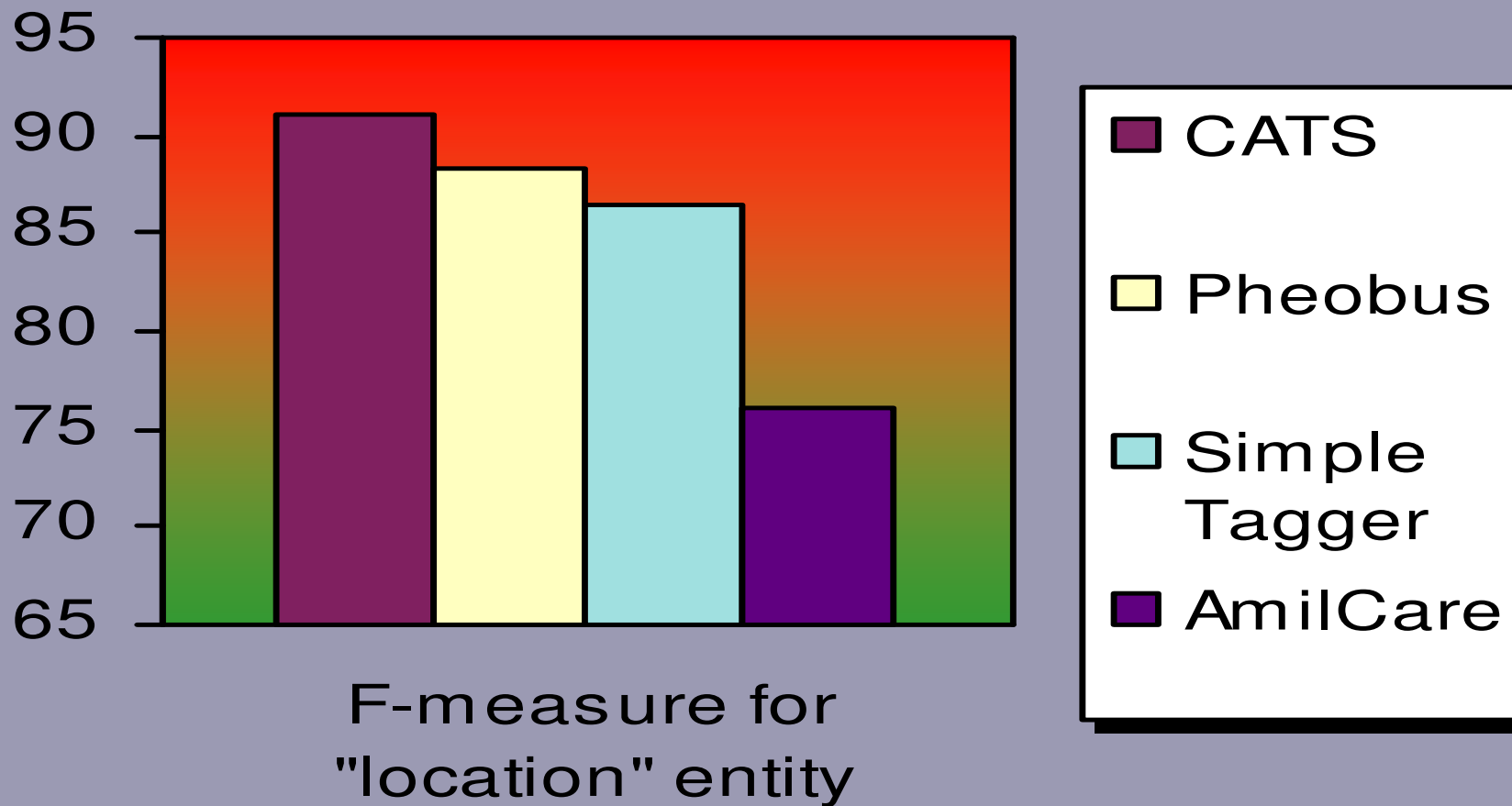
# CATS: Evaluation & Results



# CATS: comparison with other systems



# CATS: comparison with other systems



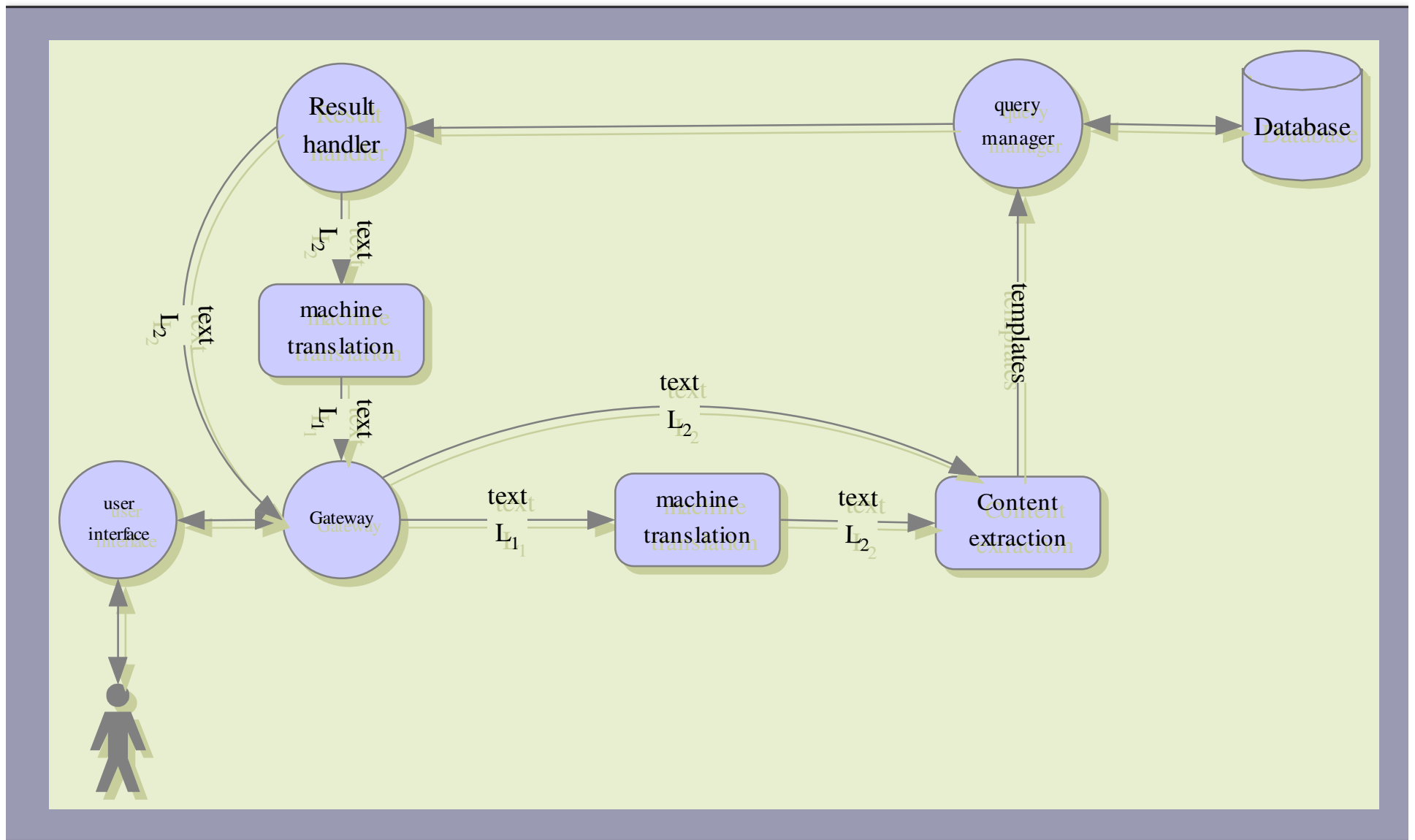
# Evaluation of CATS performance

- Average computation response time 10~30 seconds during heavy traffic (10 posts/minute)
- Compare to 7.2 seconds (CE time) using FASTUS, and 1296 seconds (CE time) using TACITUS.

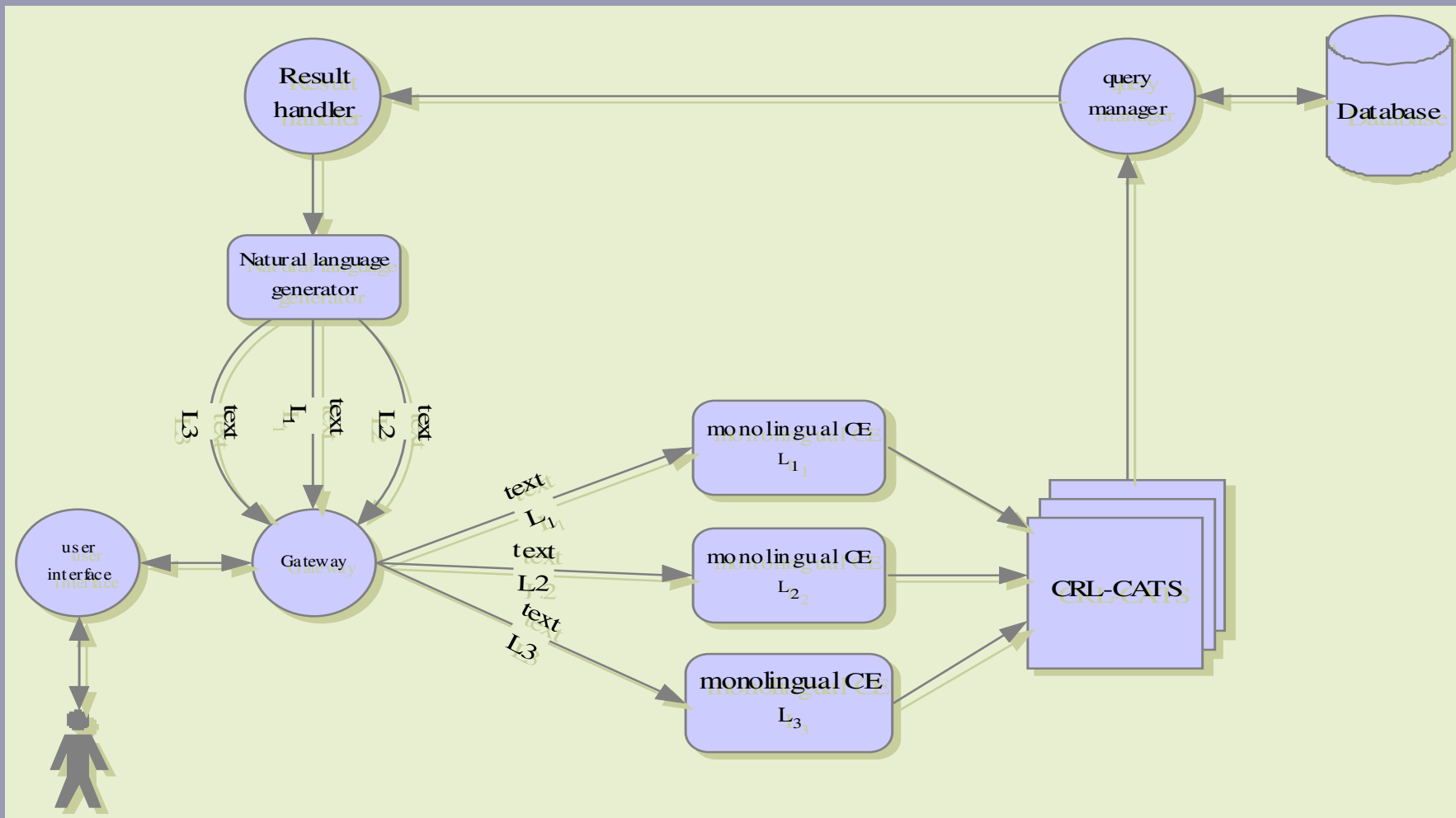
## Scalability: Adaptation to other domains

Task	resources
Collecting web-based corpus (300 posts)	2 man-hours
Analysis of the sublanguage used and modifying CRL-CATS to accommodate this domain	1 man-week
Adding domain specific entries to the dictionary (600 entries)	3 man-days
Developing the rules (100) rules is added	1 man-week
Updating the database and other related components	1 man-week (estimate)

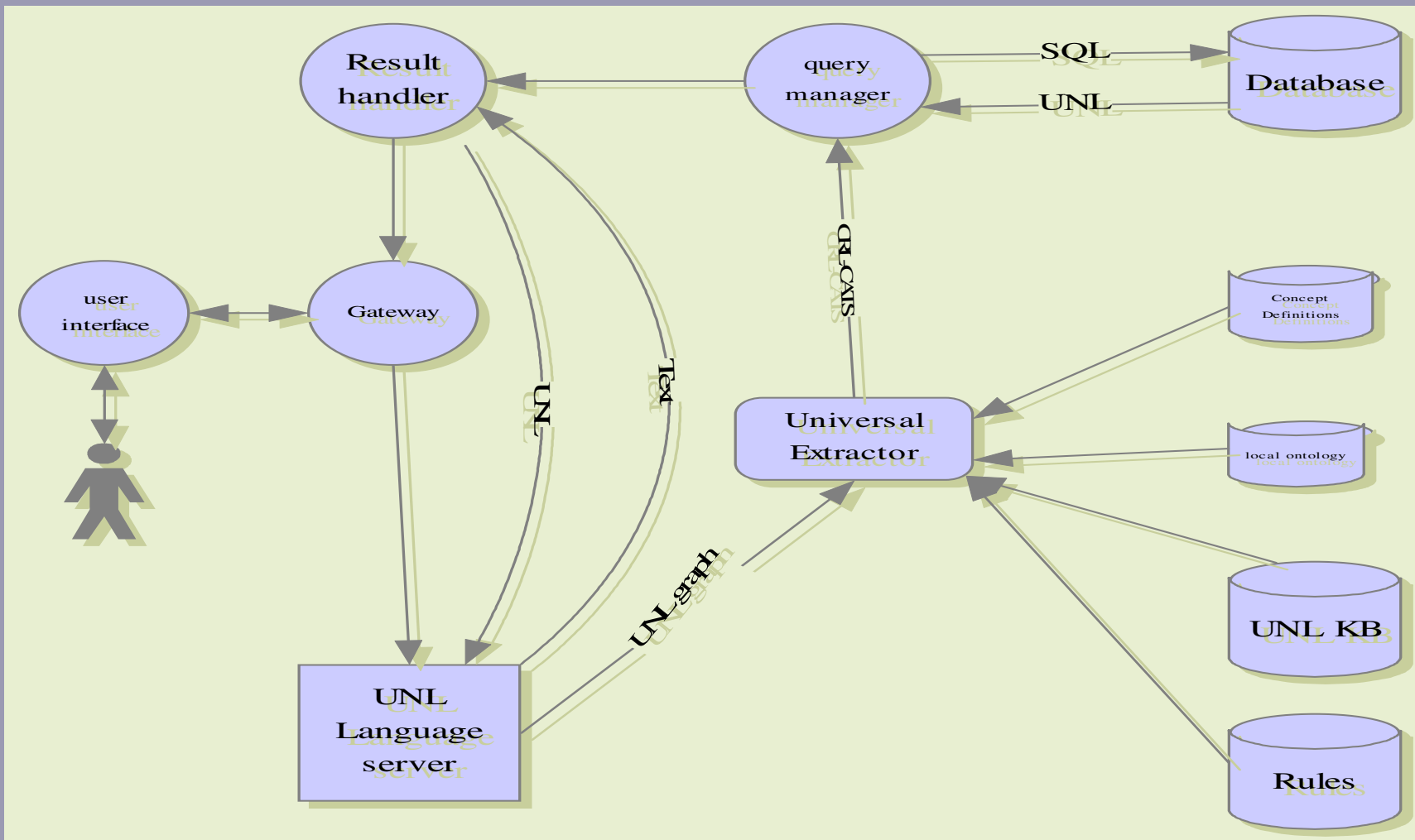
# Scalability: Extensions to other languages (MT)



# Scalability: Extensions to other languages (localization of extractors)



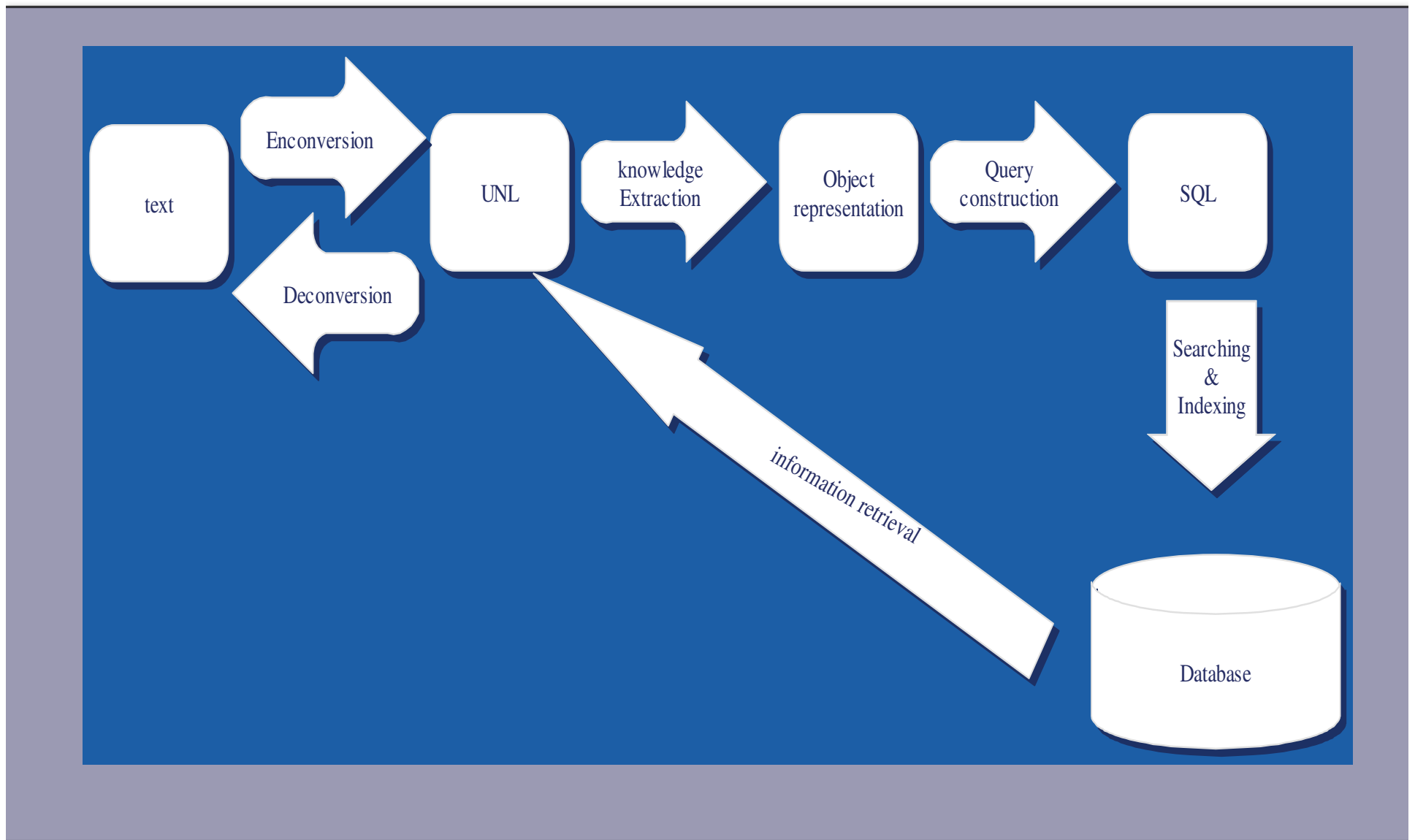
# Scalability: Extensions to other languages (the use of interlingua)



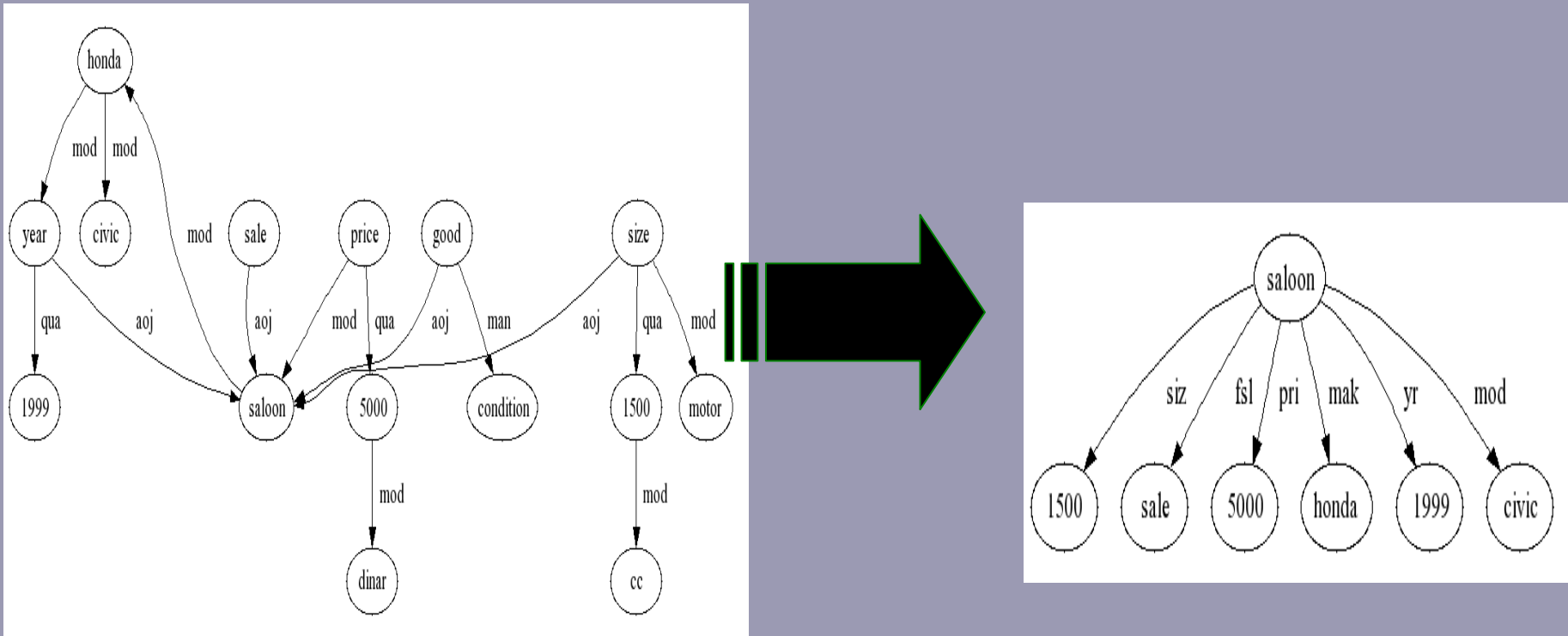
## Why do we need UE?

- The information expressed by UNL is universal.
- The same knowledge can be represented in UNL differently, making intelligent processing difficult and not accurate.
- We need a higher abstraction of knowledge.
- The UE extracts and normalizes knowledge found in UNL graphs.

# From another perspective



# object/attribute/ value representation



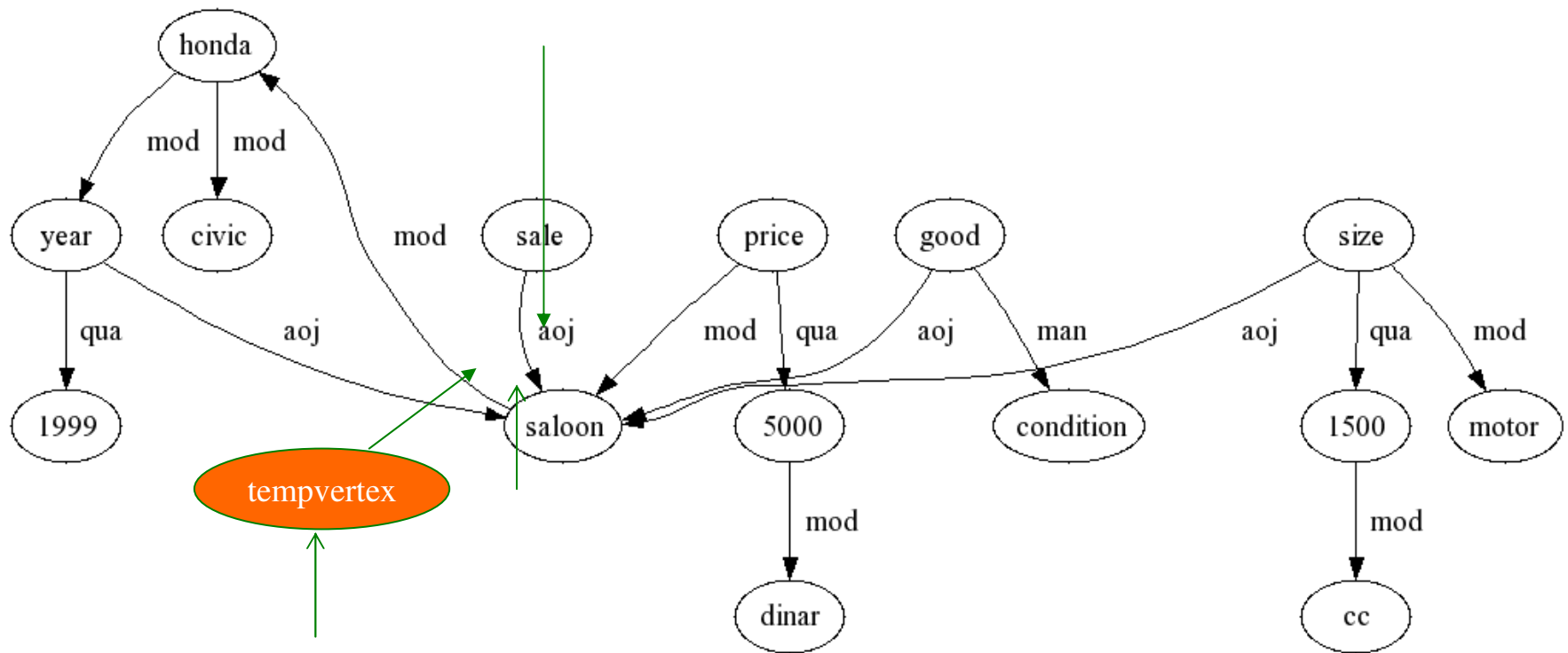
# How the UE work?

RW

```
movr:mod{tempvertex::tempedge:}{vech:+ok::}P200;
```



```
create:make{vech:+addmake:mod:}{make,~new:+new,+makeadded::}P220;
```



LW

Merci  
Thank you  
شكرا

---