A Solid Foundation of Semantic Computing toward Web Intelligence

Mitsuru Ishizuka

School of Information Science and Technology

New Tech. Committee on Semantic Computing in IEEE Computer Soc.

IEEE computer society

Search | Advanced

The world's leading membership organization for computing professionals

Publications | Conferences | Digital Library | Build Your Career | e-Learning Campus | Certification & Training | Communities | Store | JOIN

home | Mission and Vision | Join a TC | TC Chair Tools | Board | List of TCs | Conferences_List

Technical Activities

Technical Committee on Semantic Computing

The Technical Committee on Semantic Computing (SC) addresses the derivation and matching of the semantics of computational content to that of naturally expressed user intentions in order to retrieve, manage, manipulate or even create content, where "content" may be anything including video, audio, text, software, hardware, network, process, etc.

This connection between content and the user intentions is made via (1) Semantic Analys converting it to machine processable descriptions (semantics); (2) Semantic Integration, v multiple sources; (3) Semantic Applications, which utilize content and descriptions to solv which interprets users' intentions expressed in natural language or other communicative intentions of users to create content via analysis and synthesis techniques.

The ultimate success of Semantic Computing requires new, synergized technologies be d data and knowledge engineering, software engineering, computer systems and networks recognition, etc.

Founded in 2010, the mission of TCSEM is to establish a community for Semantic Computi

Resources

More Information

Join a Technical Committee, Council, or Task Force

Semantic Computing Gets Technical Committee

University of California Irvine professor Phillip C-Y Sheu, interim chair of the newly formed Technical Committee on Semantic Computing, is looking for volunteers for the Executive Committee. *—Read more*

Semantic Technology Conf.

June 2010, San Francisco

Semantic Computing

- **Toward Semantic-level Content Utilization by computers, beyond its surface-level processing.**

In many domains:

natural language texts,
image and video,
audio and speech,
semi-structured data,
behavior of software and network,
data and web mining,
etc.

Applications:

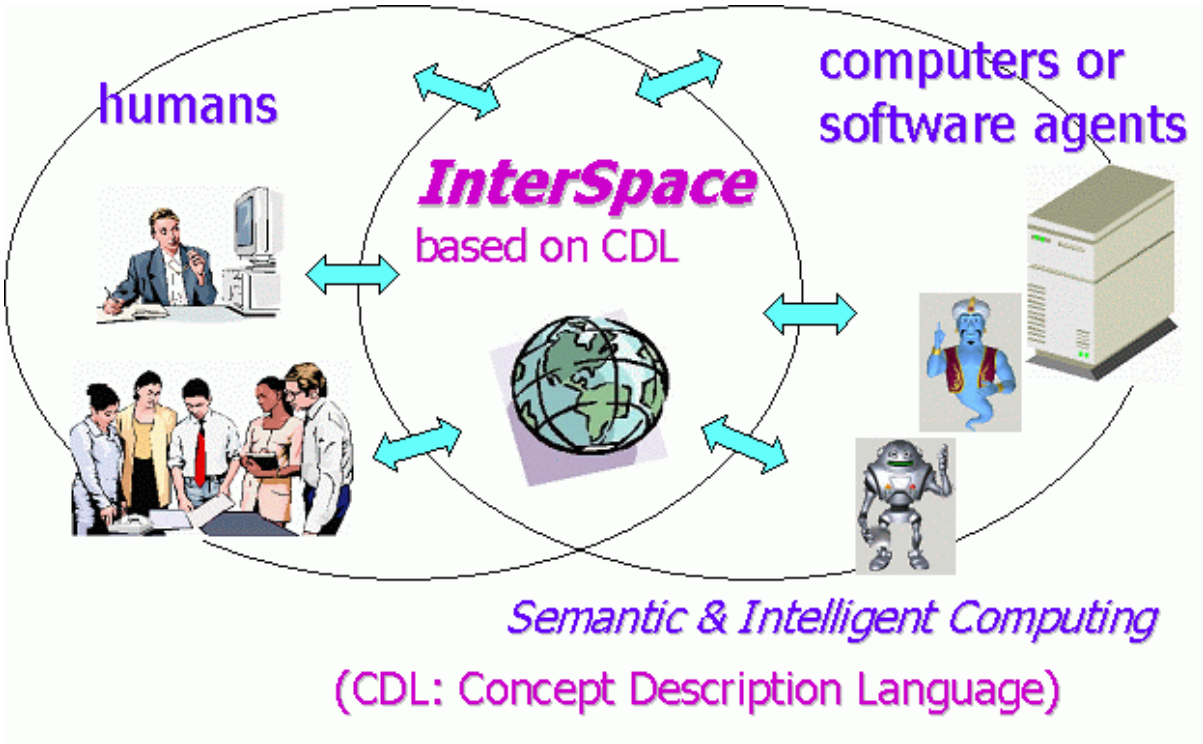
semantic annotation to contents,
semantic computing of textual documents,
semantic software engineering,
semantic search engine,
semantic multimedia services,
context-aware devices and services,
semantic GIS system,
semantic interfaces,
semantic trusted computers,
etc.

Semantic Computing at present

- Increasing interests in many domains.
- Most technologies are partial and ad hoc at present.
- We need a solid foundation of semantic computing.

- Natural language plays a major role to express and convey the semantic meaning. It should thus becomes the first focus and the core of the semantic computing
- We need a common and universal language that computers and human can understand, to represent concept meaning at a certain level.

CDL(Concept Description Language) as a solid core of semantic computing



The aims of CDL are

1) to realize machine understandability of Web text contents, and

2) to overcome language barrier on the Web.

Major Differences from Semantic Web

Semantic Web

- **Target of representation: Meta-data extracted from Web contents.**
- **Domain-dependent ontologies (which cause the difficulty of wide inter-boundary usage)**
- **RDF / OWL (description logic is hard for ordinary people to understand)**

Tim Berners-Lee says that:
"Data Web" or "Linked Data" is more adequate rather than "the Semantic Web". (2007)

Semantic Computing based on CDL

- **Target of representation: Semantic concepts expressed in texts.**
- **Universal vocabulary (+ additional specific vocabulary in a domain if necessary), and pre-defined relation set.**
- **CDL.nl (richer than RDF)**

Main body:
Institute of Semantic Computing (ISeC) in Japan
Int'l Standardization Activity:
W3C Common Web Language(CWL)-XG

Incubator Group Activity at W3C from Oct. 2006 to May 2008

W3C Incubator Activity

Common Web Language Incubator Group Charter

The **mission** of the Common Web Language Incubator Group, part of the Incubator Activity, is to develop the CWL (Common Web Language), a common language for exchanging information through the web and also for enabling computers to process information semantically. A pilot model of CWL platform will also be developed on the web. The CWL is used to describe contents and meta-data of web pages written in natural languages. Then CWL will realize a language barrier free world in the web and will also enable computers to extract semantic information and knowledge from web pages accurately.

Join the Common Web Language Incubator Group.

End date	15 October 2007
Confidentiality	Proceedings are public
Initial Chairs	Hiroshi Uchida (ISeC)
Initiating Members	• Institute of Semantic Computing (ISeC) • National Institute of Advanced Industrial Science and Technology (AIST) • Justsystem Corporation
Usual Meeting Schedule	Teleconferences: once/month Face-to-face: twice a year W3C10Asia; AC Meeting(Tokyo) 28Nov. 2006 in Japan and May 2007 in Canada

- Scope
- Deliverables
- Dependencies
- Participation
- Communication
- Decision Policy
- Patent Policy
- About this Charter

Scope

An attempt to describe texts in the web in a common language is promoted in the Semantic Web Activity. The RDF/OWL is used as a basic description language and can be used to describe texts in web pages. However, RDF/OWL is originally designed to describe meta-data of resources, and at this moment, there is no standard set of properties and vocabulary to cover various web pages. There are some activities to provide common bases for describing information in the web such as the WordNet, NICT-EDR Electronic Dictionary for providing lexical bases, Conceptual Graphs for providing a representation basis. The CWL initiative is an activity quite different from those activities. The CWL will provide not only representation scheme but also a vocabulary with semantic background. It is an initiative to integrate existing and ongoing activities for providing a common description language with unambiguous grammar and enough amount of lexicons based on the CDL (Concept Description Language) scheme aiming at describing every kind of information understandable for computers.

The CWL has the following characteristics.

1. CWL is designed to be independent from any natural languages and shall enables users to develop conversion systems between CWL and

2nd Incubator Group at W3C from June 2008

W3C Incubator Activity

Common Web Language Evaluation and Installation Incubator Group Charter

The **mission** of the Common Web Language Evaluation and Installation Incubator Group, part of the Incubator Activity, is to substantiate the CWL (Common Web Language) in actual web environment using the pilot model of the CWL platform. The CWL is a graphic language of semantic network with hyper node and is used to describe contents and meta-data of web pages in three different type of form such as UNL, CDL and RDF. The CWL platform allows people to input CWL using natural languages and display information written in CWL in natural languages. Using this CWL platform, the CWL will be evaluated from multilingualism, semantic computing and semantic web points of view. Based on these evaluation and feedback, the CWL and its platform will be bearable in actual use in the web.

Join the Common Web Language Evaluation and Installation Incubator Group.

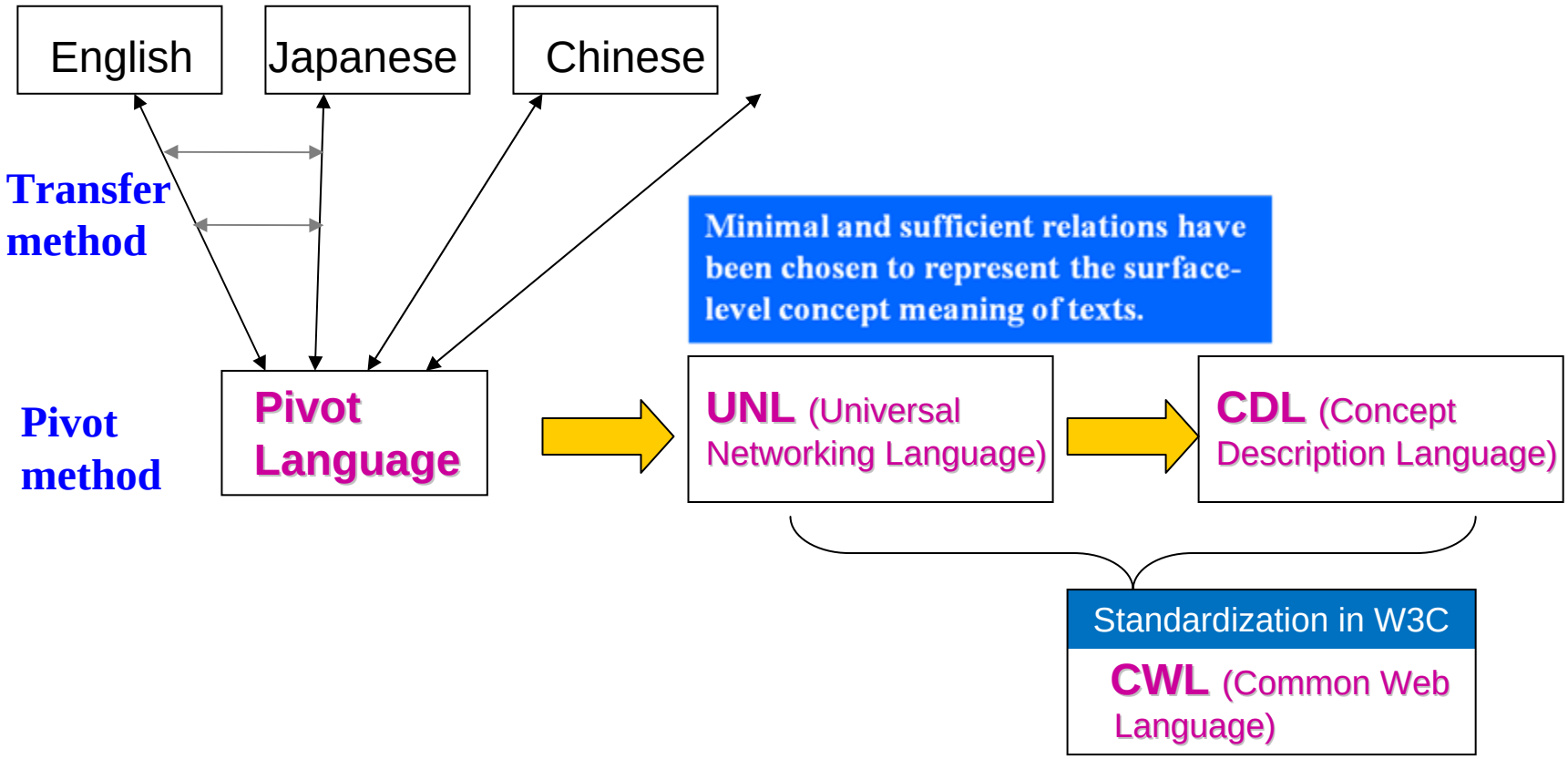
End date	20 May 2009
Confidentiality	Proceedings are public
Initial Chairs	Hiroshi Uchida (ISeC)
Initiating Members	• Institute of Semantic Computing (ISeC) • National Institute of Advanced Industrial Science and Technology (AIST) • Keio University • JustSystems Corporation
Usual Meeting Schedule	Teleconferences: No teleconferences Face-to-face: 2 per year

Scope

The CWL is a graphic language of semantic network with hyper node, a node represents a concept, an arc represents a relation between nodes and a node can be annotated by attributes. This CWL can be expressed in three forms such as UNL, CDL and RDF. The same information in CWL can be described in each form but in different manner. The CWL.unl is a language in UNL form, the CWL.cdl is a language in CDL form and the CWL.rdf is a language in RDF form. Information in the web is basically expressed in natural languages. UNL is for multilingual activities. CDL is for semantic computing activity. RDF is for semantic web activities. Various information will be expressed in three types of representation, and applications based on those representations will be developed, and information will be utilized.

Three different types of representations of CWL allow different way of treatment for the same information described. CWL.unl is unstructured text for

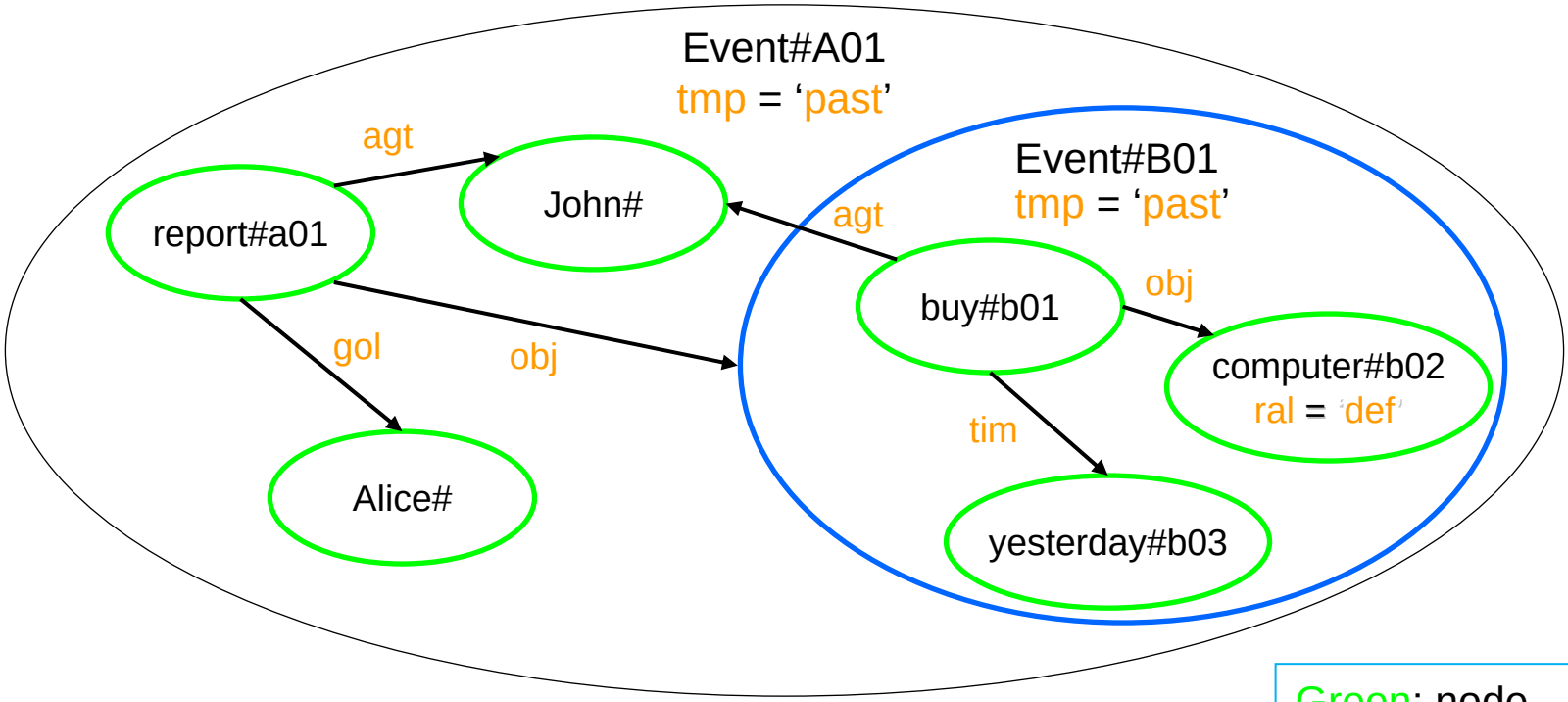
From Machine Translation



CDL Representation

- Text example:
  "John reported to Alice that he bought a computer yesterday."

- CDL graph notation:



Green: node
Blue: hyper-node

CDL Representation

- Text example:

  “John reported to Alice that he bought a computer yesterday.”

- CDL text notation:

```
{#A01 Event tmp=‘past’;
        {#B01 Event tmp=‘past’;
                <#b01:buy;>
                <#b02:computer ral=‘def’;>
                <#b03:yesterday;>
                [#b01 agt #John]
                [#b01 obj #b02]
                [#b01 tim #b03]
        }
        <#John:John;>
        <#Alice:Alice;>
        <#a01:report;>
        [#a01 agt #John]
        [#a01 gol #Alice]
        [#a01 obj #B01]
}
```

Orange: entity
Blue: relation

CDL (UNL) Relations – 44 labels

Semantic Roles

Intra-Event	
[Agent Relations]	**[Instrument Relations]**
agt (agent)	ins (instrument)
cag (co-agent)	met (method, means)
aoj (thing w/ attribute)	**[State Relations]**
cao (co-thing w/ attribute)	src (source, initial state)
ptn (partner)	gol (goal, final state)
[Object Relations]	via (interm. place or state)
obj (affected thing)	**[Time Relations]**
cob (affected co-thing)	tim (time)
opl (affected place)	tmf (initial time)
ben (beneficiary)	tmt (final time)
[Place Relations]	dur (duration)
plc (place)	**[Manner Relations]**
plf (initial place)	man (manner)
plt (final place)	bas (basis for a standard)
scn (scene)	

Logical

Inter-Entity
[Logical Relations]
and (conjunction)
orr (disjunction, alternative)
[Concept Relations]
equ (equivalent)
icl (included)
iof (an instance of)

Intra- and Inter-Event
[Cause Relations]
con (condition)
pur (purpose, objective)
rsn (reason)
[Sequence Relations]
coo (co-occurence)
seq (sequence)

Restrictive

Restrictive
cnt (content, namely)
fmt (range, from-to)
fmr (origin)
mod (modification)
nam (name)
per (proportion, rate)
pof (part of)
pos (possessor)
qua (quantity)
tto (destination)

Discourse

Semantic Role Labels in PropBank

The focus is on Predicate-Argument Structure.

- **Arg0** (prototypical agent)
- **Arg1** (prototypical patient)
- **Arg2** (indirect object/benefactive/instrument/attribute/end state)
- **Arg3** (start point/benefactive/instrument/attribute)
- **Arg4** (end point)
- **Arg5** ( )
- **TMP** (time)
- **LOC** (location)
- **DIR** (direction)
- **MNR** (manner)
- **PRP** (purpose)
- **CAU** (cause)
- **MOD** (modal verb)
- **NEG** (negative marker)
- **ADV** (general-purpose modifier)
- **DIS** (discourse particle and clause)
- **PRD** (secondary predication)

These are defined wrt each word sense.

Ex) **buy**::

Arg0: buyer
Arg1: thing bought
Arg2: seller (bought-from)
Arg3: price paid
Arg4: benefactive (bought-for)

This set is not sufficient for representing every concept expressed in natural language texts. It cannot be used for every language due to its language (English) dependency.

Rich Attributes in UNL and CDL

- **Express subjectivity evaluation of the writer/speaker for the sentence.**
- **Ex.) tense, aspect, mood, etc.**

- Time with respect to writer
  @past @present @future
- Writer's view on aspect of event
  @begin @complete @continue @custom @end @experience @progress @repeat @state
- Writer's view of reference
  @generic @def @indef @not @ordinal
- Writer's view of emphasis, focus and topic
  @emphasis @entry @qfocus @theme @title @topic
- Writer's attitudes
  @affirmative @confirmation @exclamation @imperative @interrogative @invitation @politeness @respect @vocative
- Writer's view of reference
  @generic @def @indef @not @ordinal

- Writer's feeling and judgements
  @ability @get-benefit @give-benefit @conclusion @consequence @sufficient @grant @grant-not @although @discontented @expectation @wish @insistence @intention @want @will @need @obligation @obligation-not @should @unavoidable @certain @inevitable @may @possible @probable @rare @regret @unreal @admire @blame @contempt @regret @surprised @troublesome
- Describing logical characters and properties of concepts
  @transitive @symmetric @identifiable @disjoint
- Modifying attribute on aspect
  @just @soon @yet @not
- Attribute for convention
  @passive @pl @angle_bracket @brace @double_parenthesis @double_quote @parenthesis @single_quote @square_bracket

The defining method of one unique sense of a word in *UW* (Patent of UN Univ.)

- **Defining category**

swallow(icl>bird)	the bird "One swallow does not make a summer"
swallow(icl>action)	the action of swallowing "at one swallow"
swallow(icl>quantity)	the quantity "take a swallow of water"

- **Defining possible case relations**

spring(agt>thing,obj>wood)	bending or dividing something
spring(agt>thing,obj>mine))	blasting something
spring(agt>thing,obj>person, src>prison))	escaping (from) prison
spring(agt>thing,gol>place)	jumping up "to spring up"
spring(agt>thing,gol>thing)	jumping on "to spring on"
spring(obj>liquid)	gushing out "to spring out"

UW (Universal Words) in UNL

Universal Word
uw{(equ>Universal Word)}
adjective concept{(icl>uw)}
uw(aoj>thing{,and>uw,ben>thing,cao>thing,cnt>uw,cob>thing,con>uw,coo>uw,dur>period,man>how,obj>thing,or>uw(aoj>thing),plc>thing,plf>thing,plt>thing,rsn>uw(aoj>thing),rsn>do,icl>adjective concept})
Achaean({icl>uw(}aoj>thing{)})
Afghan({icl>uw(}aoj>thing{)})
African({icl>uw(}aoj>thing{)})
African-American({icl>uw(}aoj>thing{)})
Ainu({icl>uw(}aoj>thing{)})
Alaskan({icl>uw(}aoj>thing{)})
Albanian({icl>uw(}aoj>thing{)})
Aleutian({icl>uw(}aoj>thing{)})
Alexandrian({icl>uw(}aoj>thing{)})
Algerian({icl>uw(}aoj>thing{)})
Altaic({icl>uw(}aoj>thing{)})
American({icl>uw(}aoj>thing{)})
Anglian({icl>uw(}aoj>thing{)})
Anglo-American({icl>uw(}aoj>thing{)})
Anglo-Catholic({icl>uw(}aoj>thing{)})
Anglo-French({icl>uw(}aoj>thing{)})
Anglo-Indian({icl>uw(}aoj>thing{)})
Anglo-Irish({icl>uw(}aoj>thing{)})
Anglo-Norman({icl>uw(}aoj>thing{)})
Arab({icl>uw(}aoj>thing{)})
Arab-Israeli({icl>uw(}aoj>thing{)})
Arabian({icl>uw(}aoj>thing{)})
Arabic({icl>uw(}aoj>thing{)})

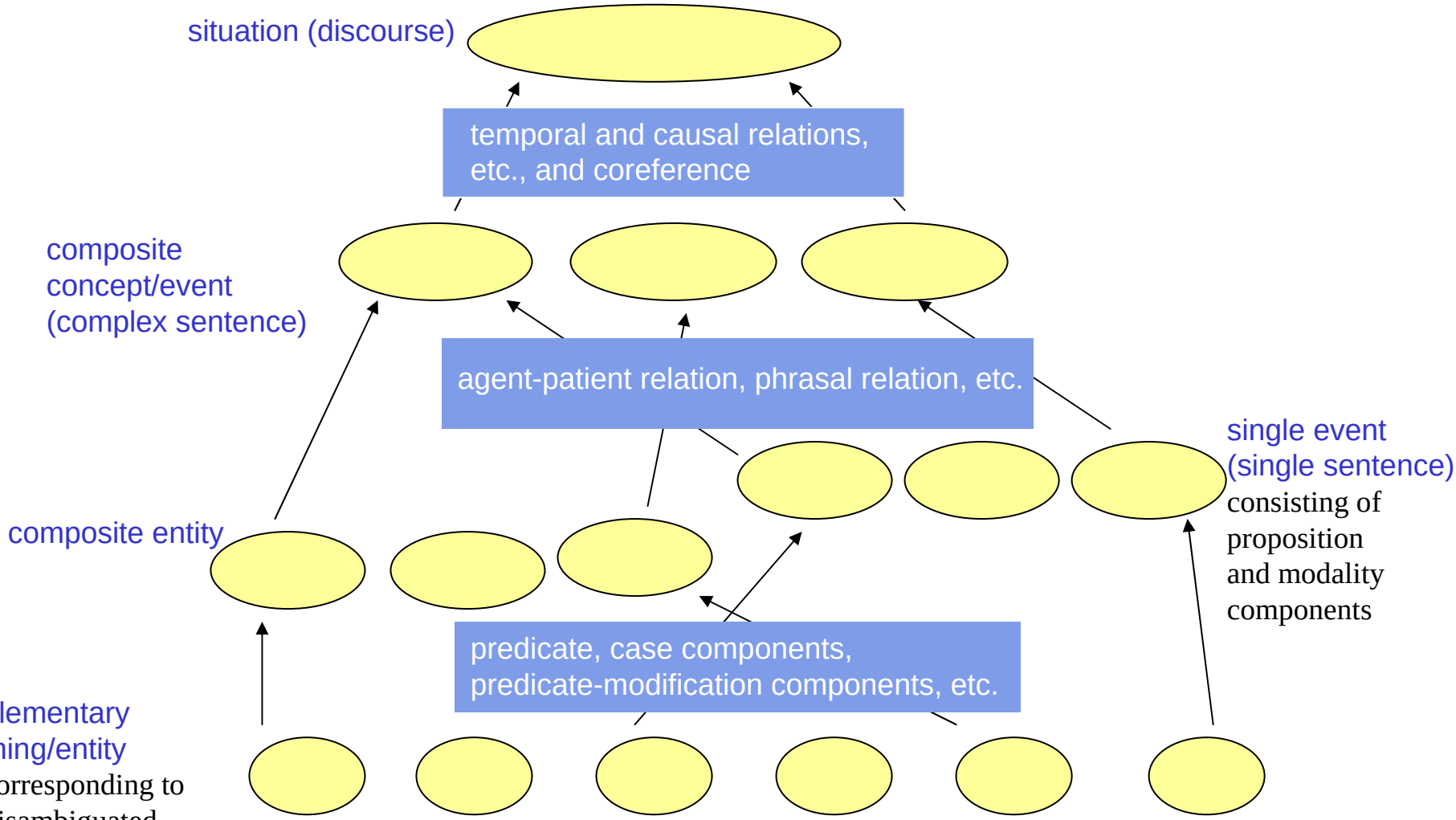
40,000 lexicons are open to public.

The full vocabulary includes 200,000 lexicons as of 2007.

Concept Description Levels

- **There are several choices for the deep semantic-level description depending on applications. On the other hand, a certain consensus has been made wrt "Concept Description" which is slightly below the surface level, through decades-long researches on NLP, machine translation and electric dictionaries.**
- **Whereas a complete consensus has not been achieved yet regarding the Concept Description level and its description scheme, it is meaningful to set up a common concept description format as an international standard today.**

Hierarchical Construction of Concept Representation in CDL

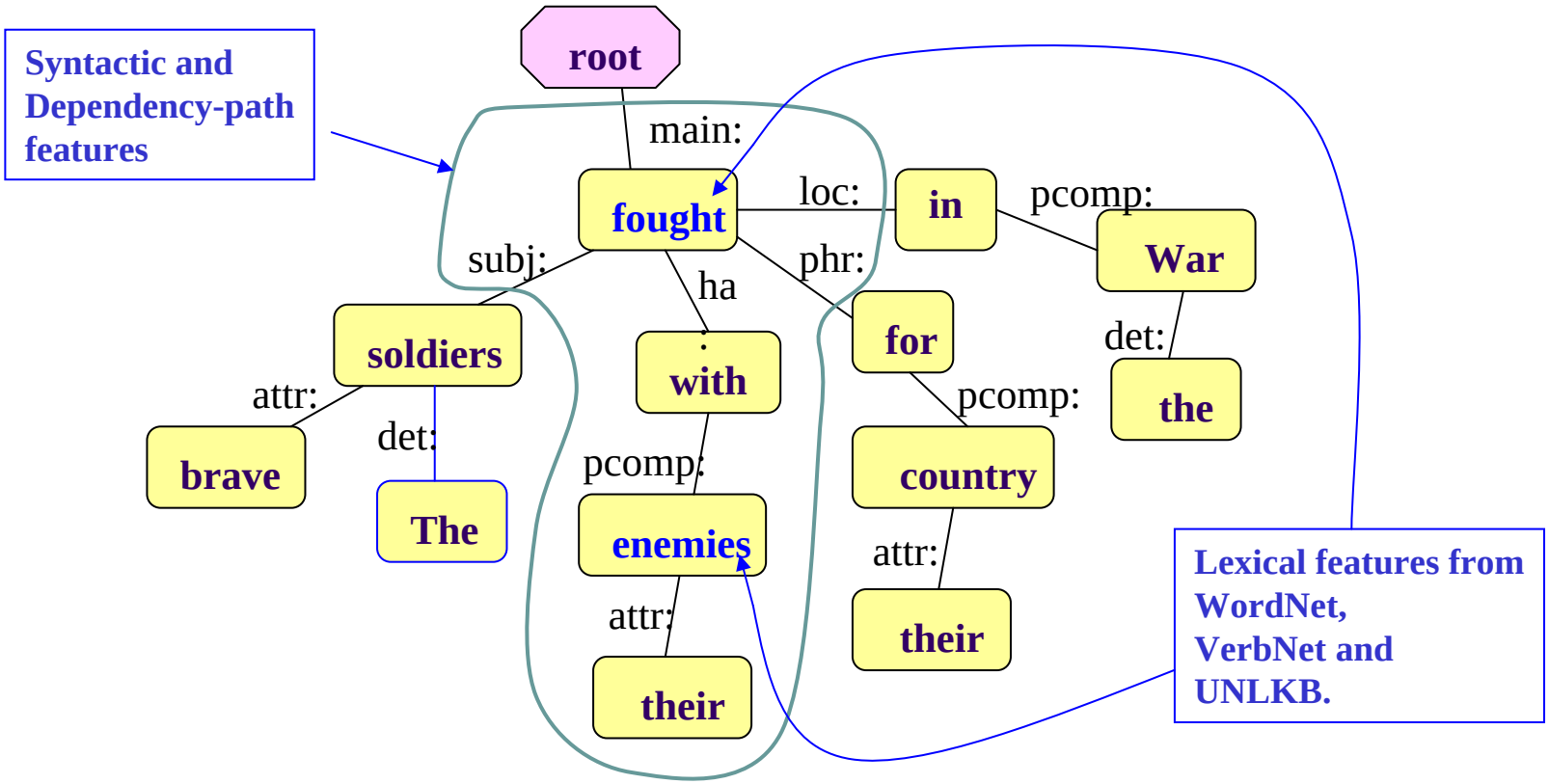


Approaches for Generating CDL Data

- **Manual Coding & Editing**
  - **Even in this case, a graphical input editor is necessary.**
- **Graphical Input & Editing** (**Hasida's Semantic Authoring**)
- **Some Manual Tagging to Text, then Conversion into CDL.**
- **Semi-automatic Conversion from Text (1)**
  - **Automatic and Manual Word Sense Disambiguation, then Conversion into CDL.**
- **Semi-automatic Conversion from Text (2)**
  - **Post editing of converted CDL data with a GUI.**
- **Full Automatic Conversion** (**ultimate goal**)

Our current approach

Recognition of CDL Relations from dependency-analyzed text



Some labels of Connexor Machinese Analyser:

ha (prepositional phase attachment), **phr** (verb particle), **pcomp** (subject complement)

Performance for frequent 36 relations (out of 44)

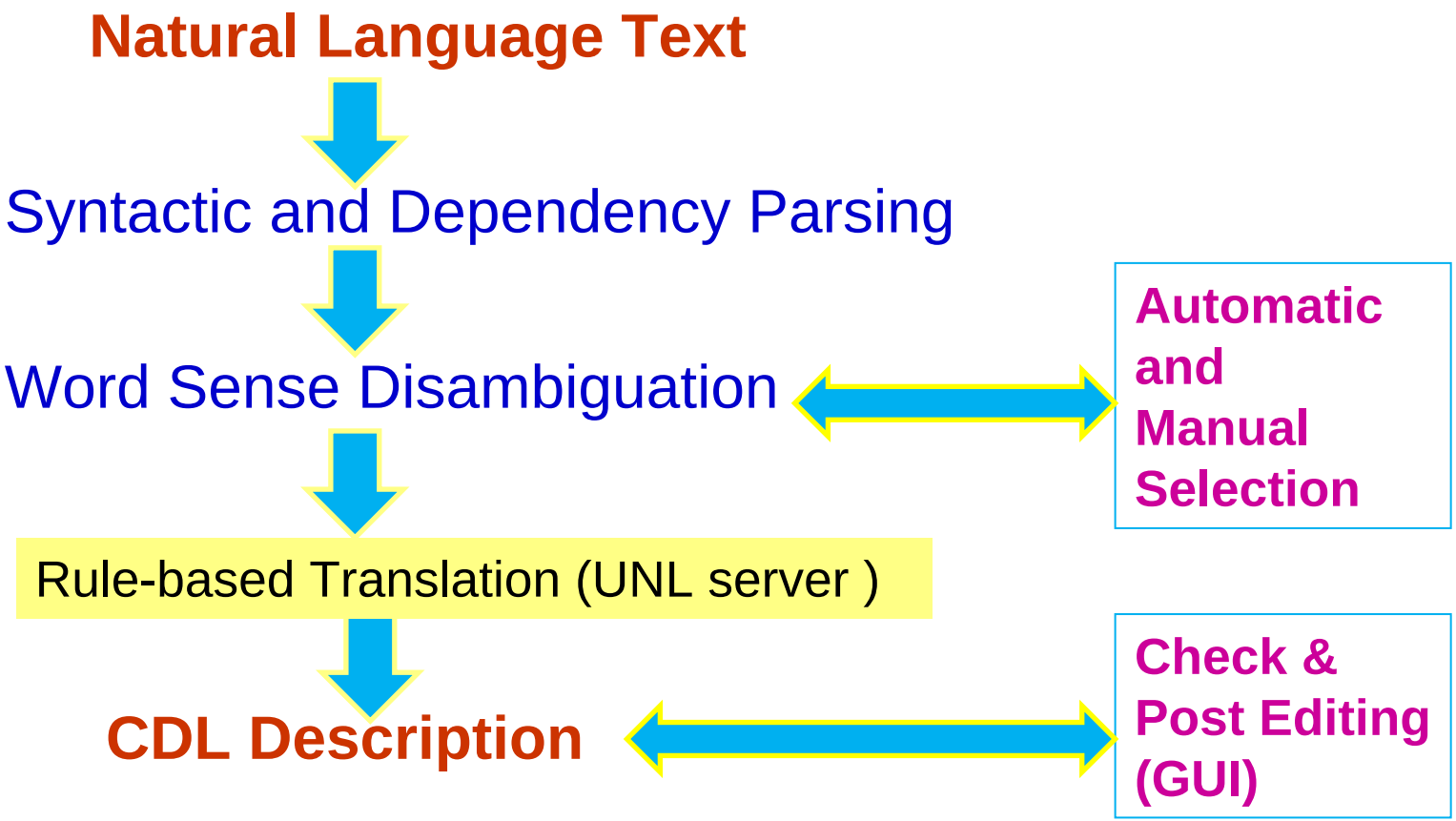
Precision 87.3% Recall 88.1% F-value 87.1%

Frequencies of CDL Relations

- **Data sparseness** :
  - The whole number of relation:13487
  - Relation type: 44
  - Average num per relation: 306.5

nam	Mod	Obj	Aoj	And	Agt	Man	Plc	Gol	Tim	Pur	Qua
#rel	3128	2697	2069	1122	1046	788	446	395	321	289	269
nam	Pos	Scn	Rsn	Src	Cnt	Dur	Bas	Met	Equ	Nam	Con
#rel	86	71	65	63	61	58	49	47	46	41	41
nam	Ben	Tmt	Pof	Frm	Or	Fmt	Tmf	Seq	To	Iof	Cag
#rel	27	25	24	23	21	20	19	17	12	11	10
nam	Icl	Via	Coo	Per	Ins	Plt	Ptn	Plf	Cao	Opl	Cob
#rel	10	9	8	8	8	7	6	4	2	1	0

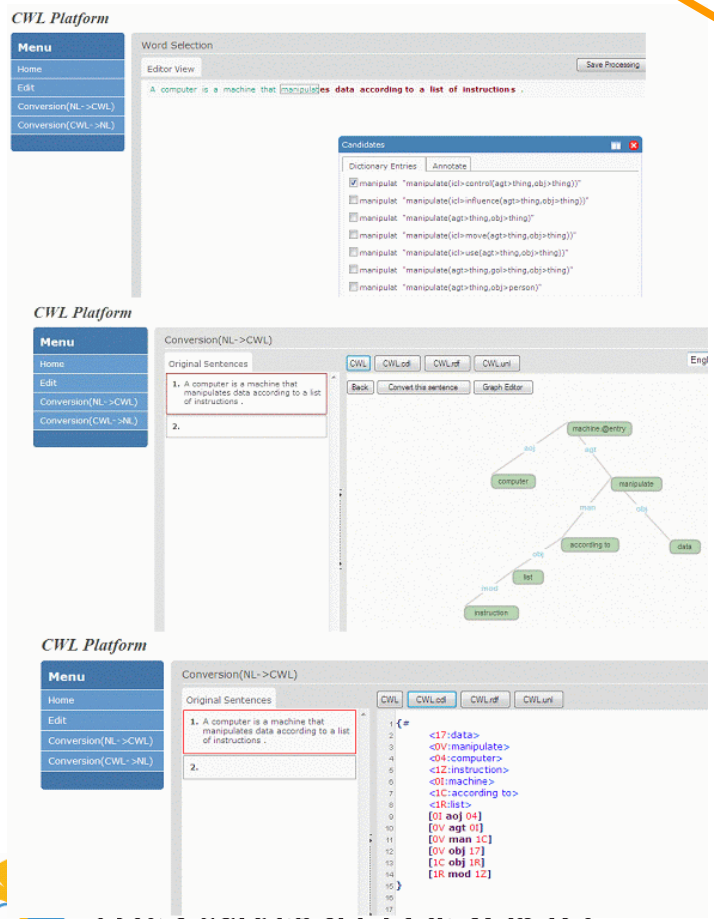
A Semi-automatic Conversion from NL Text to CDL

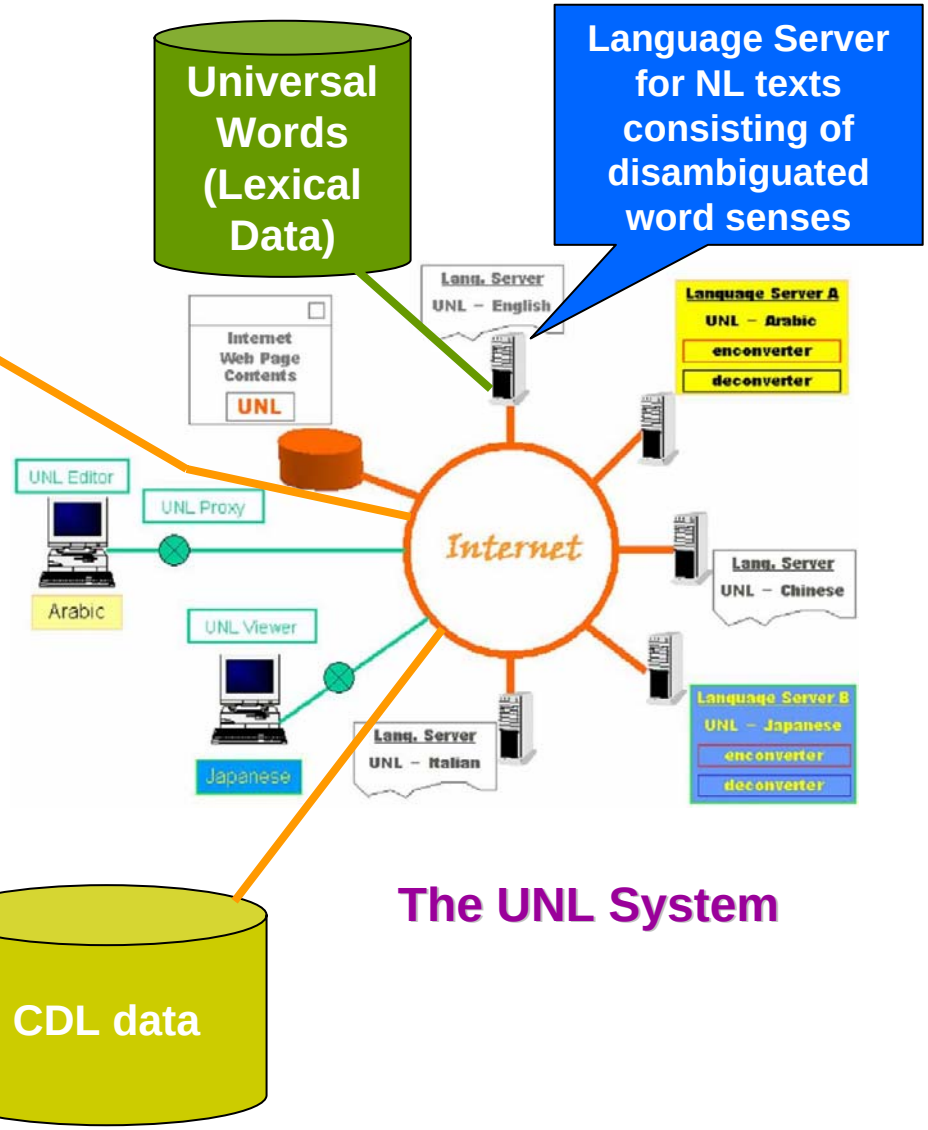


Semi-automatic Conversion from NL Texts to CDL

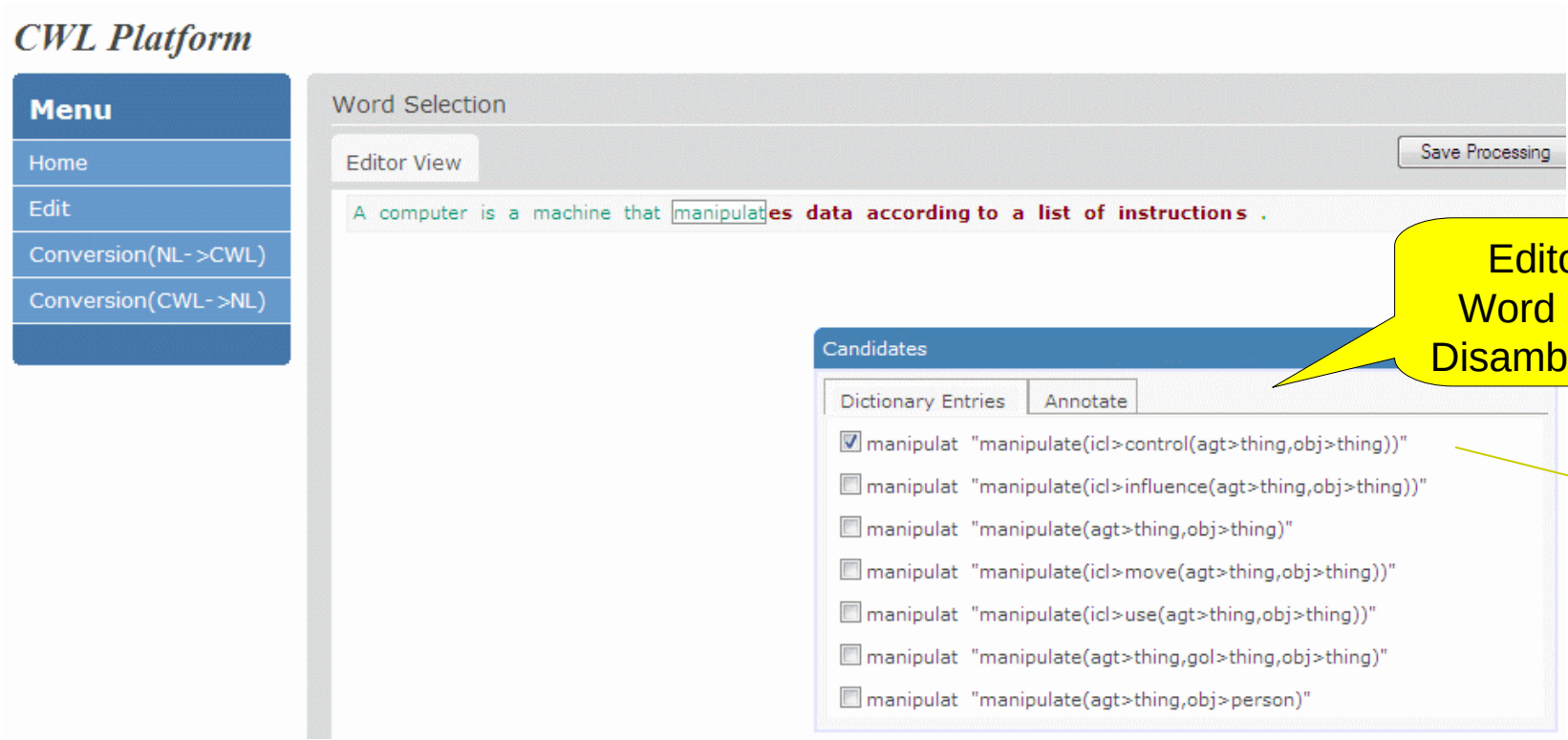
CWL Platform Interface

manual word sense disambiguation



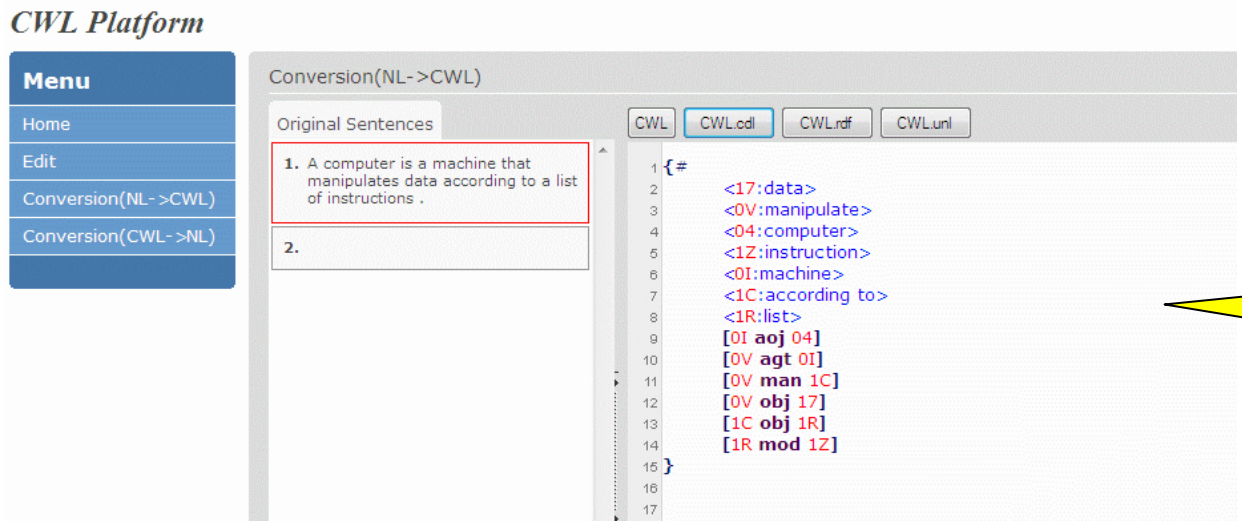


CWL Platform Interface (1)

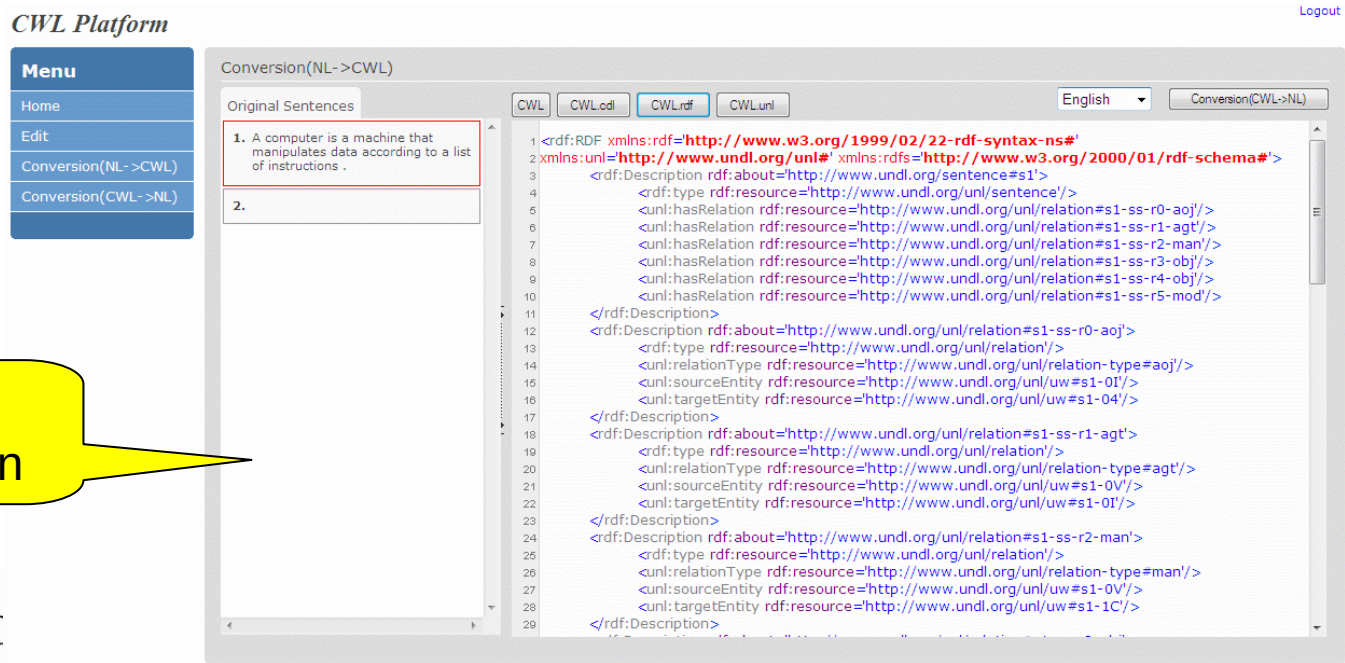


☐manipulat “manipulate(icl>control(agt>thing, obj>thing))”

CWL Platform Interface Screenshots (2)

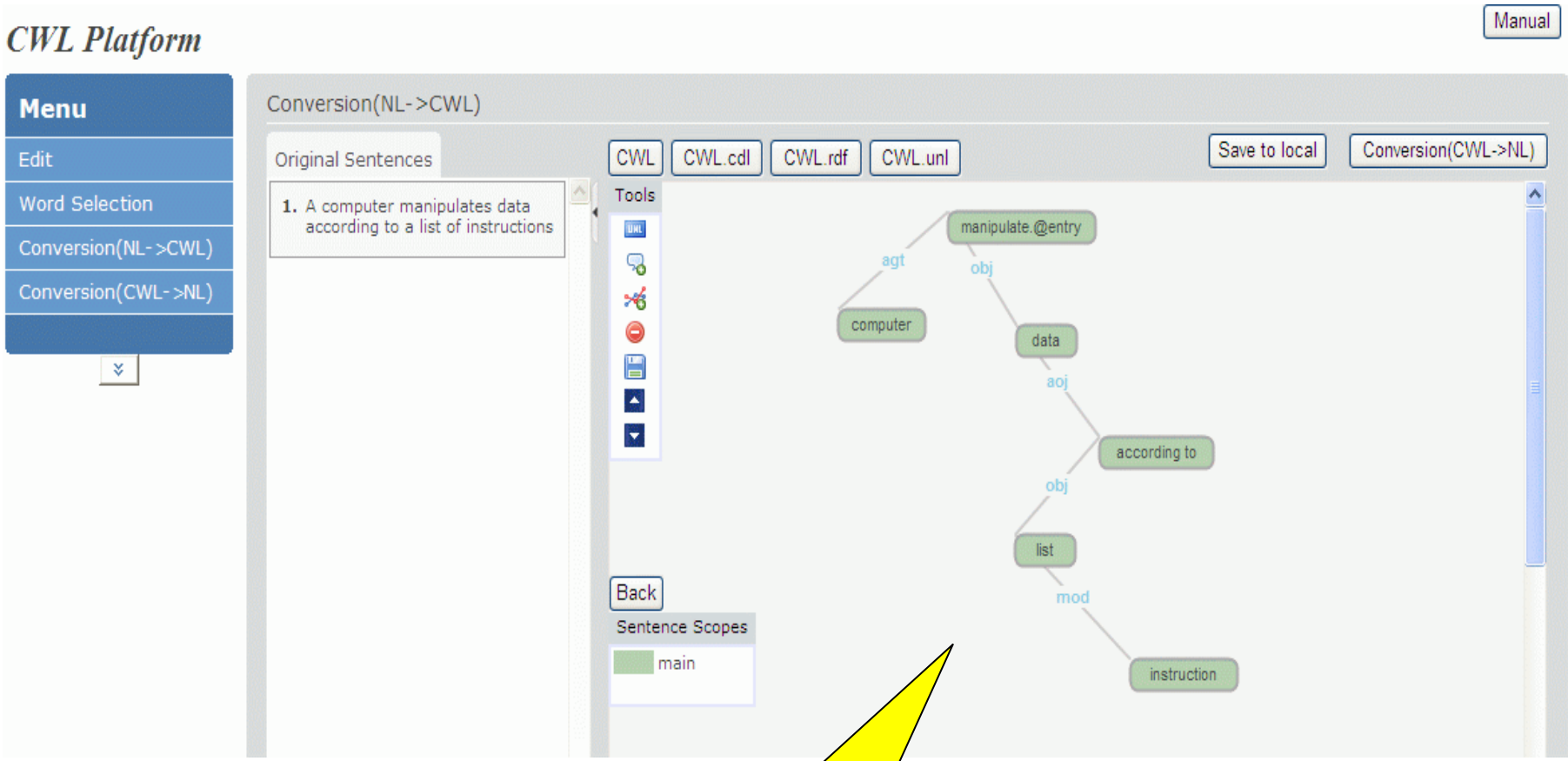


CDL description



RDF description

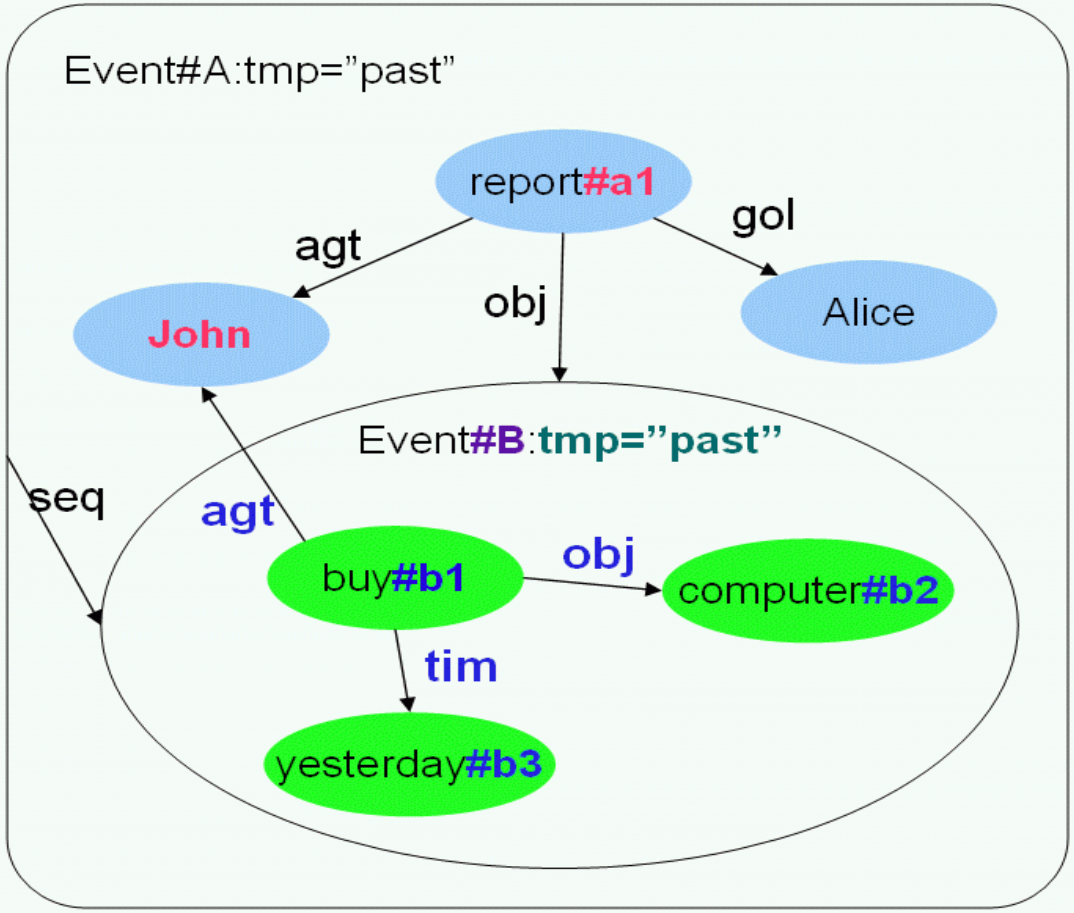
CWL Platform Interface (3)



Graph Representation

CDL Data Retrieval via CDQL

(an Extended SPARQL)



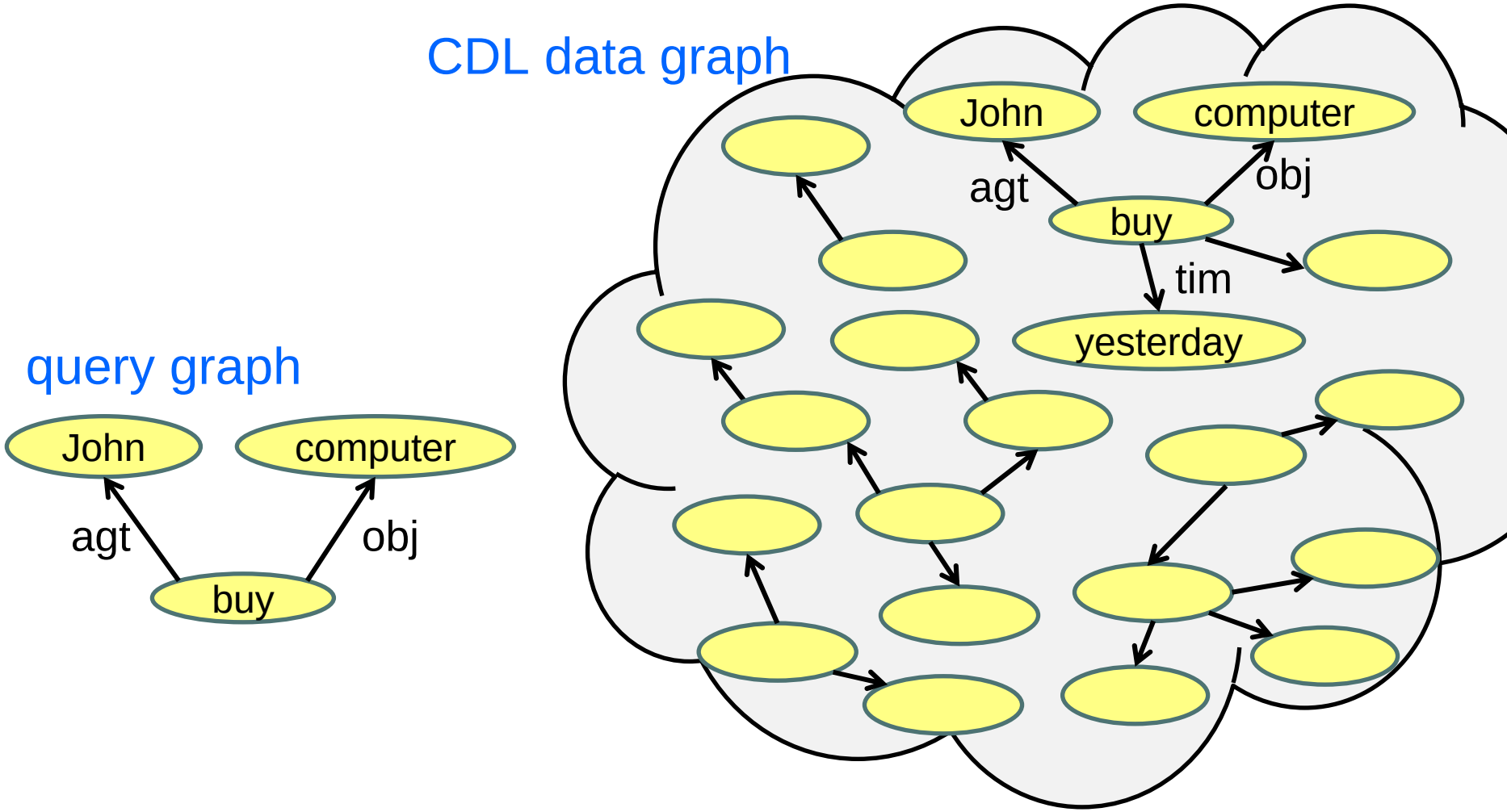
Query::
What did John report?

```
SELECT *y *z
WHERE {
  [report agt John]
  [report obj ?x]
  {?x Event; *y, *z}
}
```

result

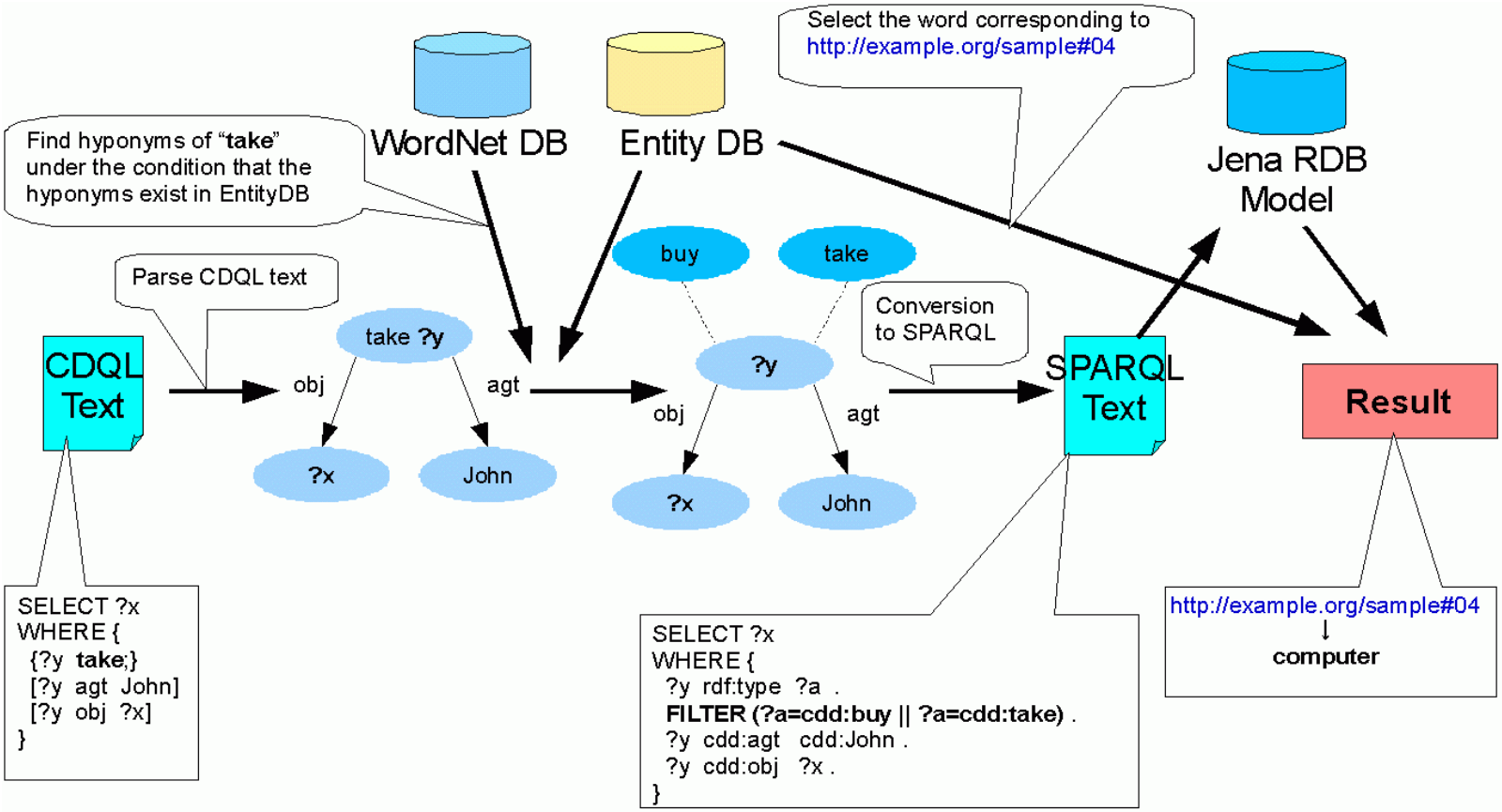
*y = {#b1 buy;}
{#b2 computer;}
{#b3 yesterday;}
*z = [#b1 agt John]
[#b1 obj #b2]
[#b1 tim #b3]

Semantic Retrieval through a Flexible Graph Matching



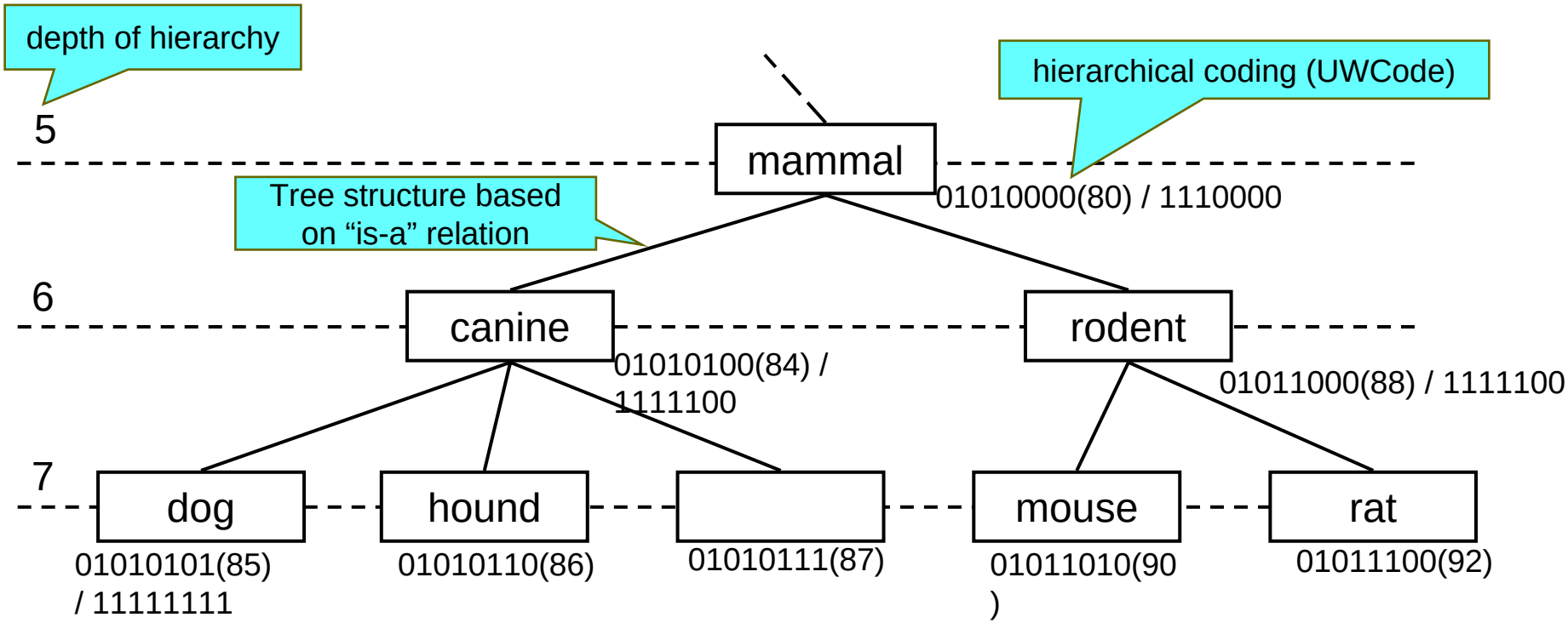
Semantic Retrieval of CDL data

- **CDQL: SQL-like query language for CDL data**

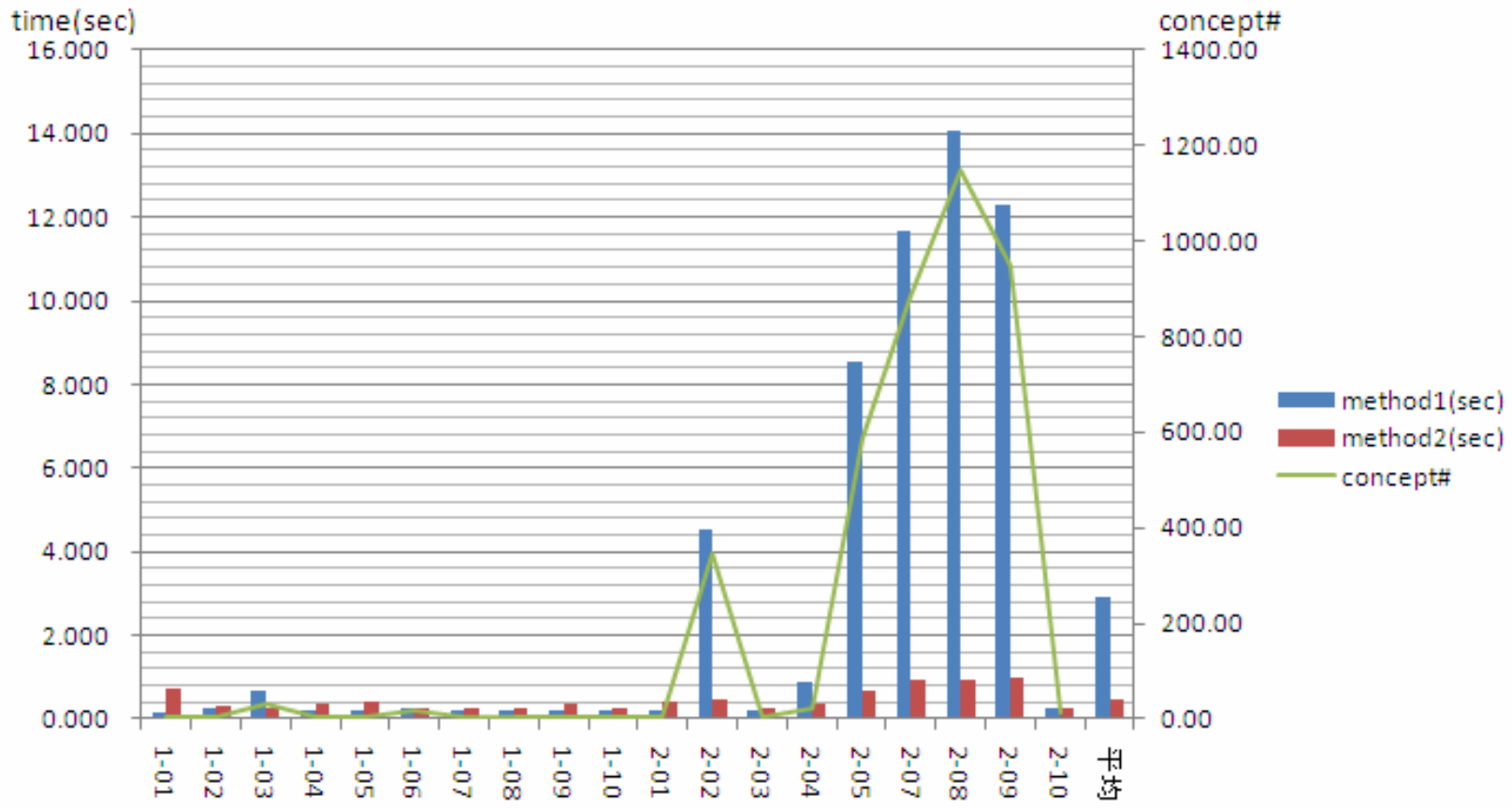


Hierarchical Coding of UW for Efficient Semantic Retrieval

- **Allow efficient controlled matching with the hyponyms, hypernyms and sibling words.**
- **64 bytes (4 bits per layer) for 20,000 words; 128 bytes for 200,000 words.**



Preliminary Result of Retrieval Speed Improvement



Summary

- **Toward a solid foundation of Semantic Computing, I introduced CDL (Concept Description Language), which is expected to be a common platform of expressing the meaning of every concept corresponding to natural language text.**
- **CDL is computer Esperanto language that both humans and computers can understand.**
- **It will also contribute to overcome the language barrier on the Web and in the world.**
- **The current major issue of CDL is a way to convert natural language texts into CDL with a small effort.**

Thank You