The Semantic Web Needs Anaphora Resolution

Rodolfo Delmonte

Dipartimento Scienze del Linguaggio

Università Ca' Foscari

Ca' Garzoni-Moro - San Marco 3417 - 30124 VENEZIA

Outline

- Introduction
- Input to the QA Module
- RDFs and Semantic Web
- Partial and Complete System
- Discourse Model
- Anaphora Resolution in Summaries

Introduction

- Question Answering and Summarization on the Web are feasible

- Following the Semantic Web initiative people use triples or ternary expressions as useful counterparts to linguistic representations

- RDFs and ternary structures are insufficient to cope with natural language texts… because of Anaphora Resolution

Semantic Web and Inferencing

- For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning.
- Meaning is expressed by RDF, which encodes it in sets of triples being rather like the subject, verb and object of an elementary sentence. These triples can be written using XML tags. In RDF, a document makes assertions that particular things (people, Web pages or whatever) have properties (such as "is a sister of", "is the author of") with certain values (another person, another Web page).

Semantic Web and Inferencing

- This structure turns out to be a natural way to describe the vast majority of the data processed by machines. Subject and Object are each identified by a URI, just as used in a link on a Web page... The verbs are also identified by URIs, which enables anyone to define a new concept, a new verb, just by defining a URI for it.

- **Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. Scientific American (May 2001).**

Semantic Web and RDFs

- The RDF data model, as specified in RDFMS defines a simple model for describing interrelationships among resources in terms of named properties and values.
- The RDF Schema mechanism provides a basic Type System for use in RDF models
- The schema specification language is a declarative representation language influenced by ideas from KR etc.

Ternary Expressions

- Ternary expressions(T-expressions), <subject relation object>.
  Certain other parameters (adjectives, possessive nouns, prepositional phrases, etc.) are used to create additional T-expressions in which prepositions and several special words may serve as relations.
  For instance, the following simple sentence

  (1) Bill surprised Hillary with his answer

  will produce two T-expressions:

  (2) <<Bill surprise Hillary> with answer>

  <answer related-to Bill>

Triples at CL

Kenneth C. Litkowski, Syntactic Clues and Lexical Resources in Question-Answering

- The key step in the CL Research question-answering prototype was the analysis of the parse tree to extract semantic relation triples and populate the databases used to answer the questions

- A semantic relation triple consists of a discourse entity, a semantic relation which characterizes the entity's role in the sentence, and a governing word to which the entity stands in the semantic relation.

Semantic Relations in Triples

- The semantic relations in which entities participate are intended to capture the semantic roles of the entities, as generally understood in linguistics.
- This includes such roles as Agent, Theme, Location, Manner, Modifier, Purpose, and Time
- Surrogate place holders include SUBJ, OBJ, TIME, NUM, ADJMOD, and the prepositions heading prepositional phrases

Grammatical Relations and Governing Predicate

- For SUBJ, OBJ and TIME this is the main verb of the sentence.
- For prepositions, it is generally the noun or verb that the preposition modified.
- For the adjectives and numbers it is the noun that is modified.

Arguments Reversibility, but not only that…

- **The IR/IE BOWs approach suffers (at least) from Reversible Arguments Problem (Katz & Lin)**
  - What do frogs eat? vs
  What eats frogs?

  -The president of Russia visited the president of China. Who visited the president?

SURFACE CONSTITUENCY RELATIONS

- **John killed Tom. Tom was killed by a man. Who killed the man?**

Problematic structures for BOWs and Ternary Expressions

- **Subject vs Object**
  - **Passivized structures**
  - **Inchoativized structures**
  - **Ergativized structures**
- **Control in Open Predicative Structure**
  - **Relative Clauses, Adjectival Adjuncts**
  - **Infinitives, Participials, etc.**

Complete System pipeline

Level One takes care of the Sentential Level Analysis in broad terms

Complete System pipeline

Does anaphora resolution at sentence level and binds all syntactic and functional control relations, i.e. relative and interrogative clauses, infinitives and participials etc.

Complete System pipeline

Level 2 works at Discourse Level

Produces a complete semantic interpretation

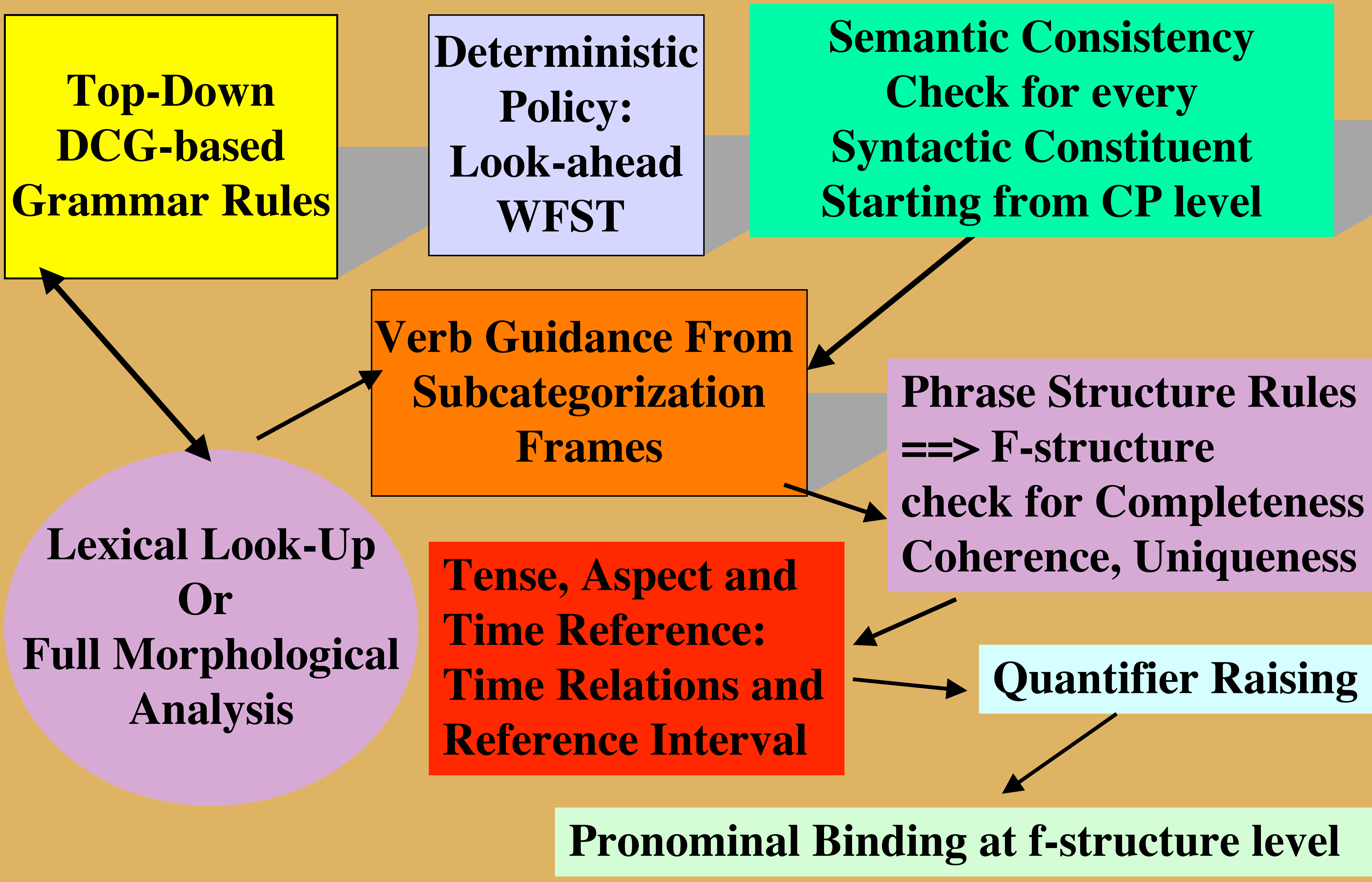
Complete System pipeline

Takes care of Topic Hierarchy and Anaphora Resolution□

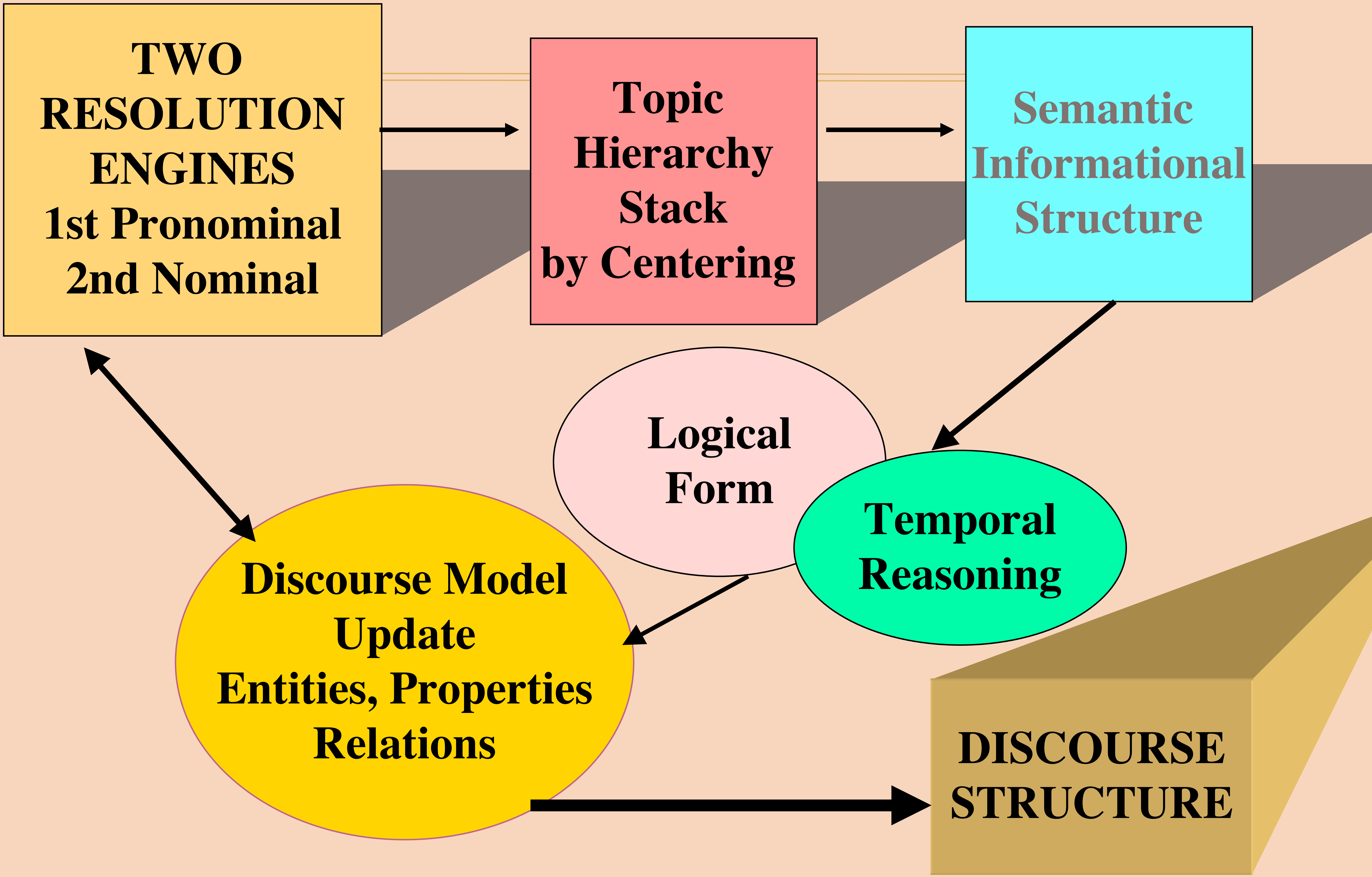
Complete System pipeline

Does semantic mapping and takes care of rhetorical structure information, builds the complete semantic interpretation and the Discourse Model. In a final process, Discourse Structure is built.

SYSTEM ARCHITECTURE I°



SYSTEM ARCHITECTURE II°



SHALLOW & COMPLETE

Complete → **Complete Parsing & Semantics Deep Anaphora Resolution**

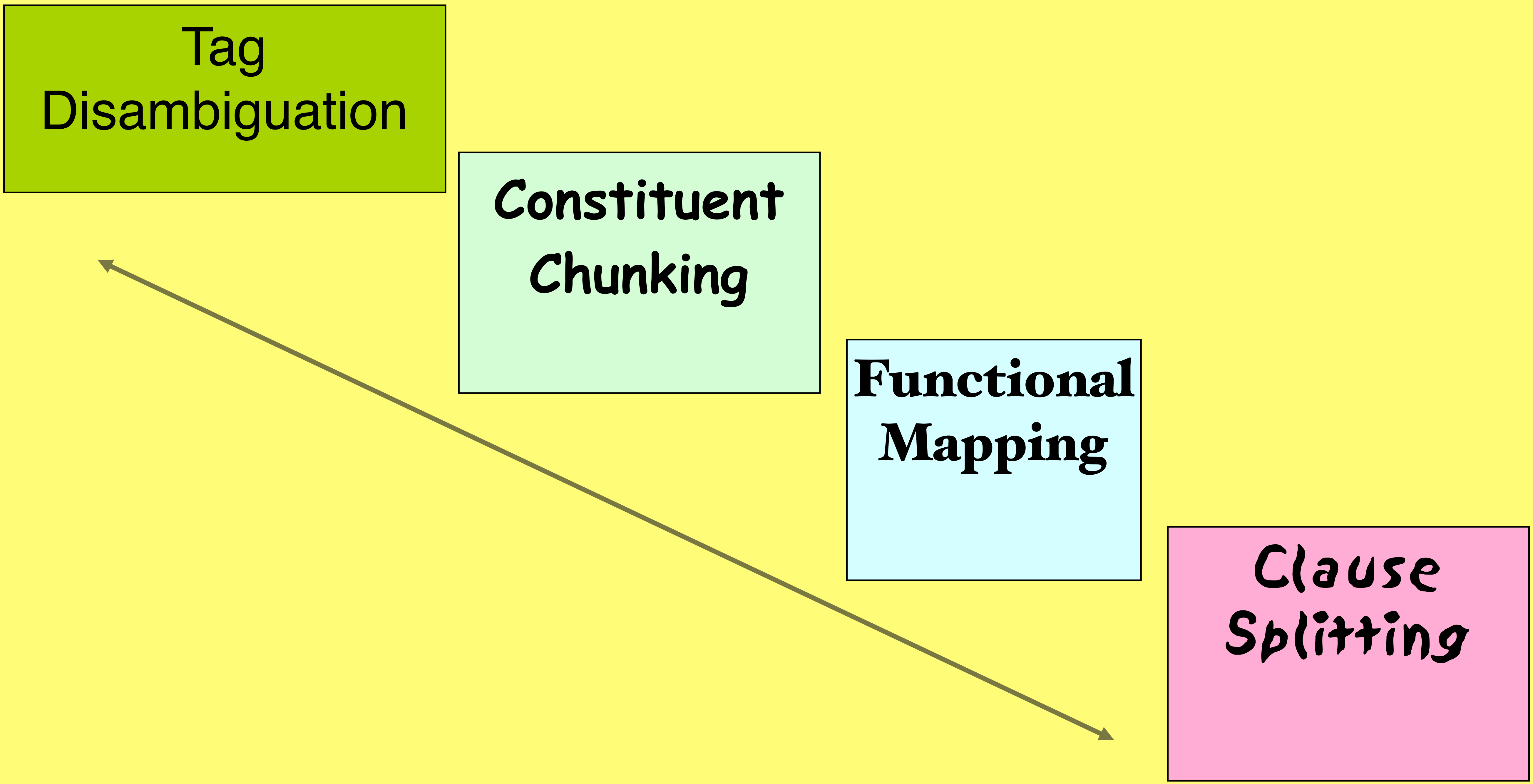
Partial

Robust

Robust & Partial Parsing... Semantics... Anaphora Resolution

Chunks → **Robust Parsing… No Semantics at Propositional Level… Shallow Anaphora Resolution**

ROBUST SYSTEM PIPELINE



Hard to realize tasks in a robust system

- Tag disambiguation
- Recognition of clausal structure
- Recognition of arguments from adjuncts
- Recognition of predicate-argument structures
- Anaphora resolution

Robust Parsing Techniques: Coping with Uncertainty

- Tag Disambiguation

95%

- Sentence Splitting into Clauses

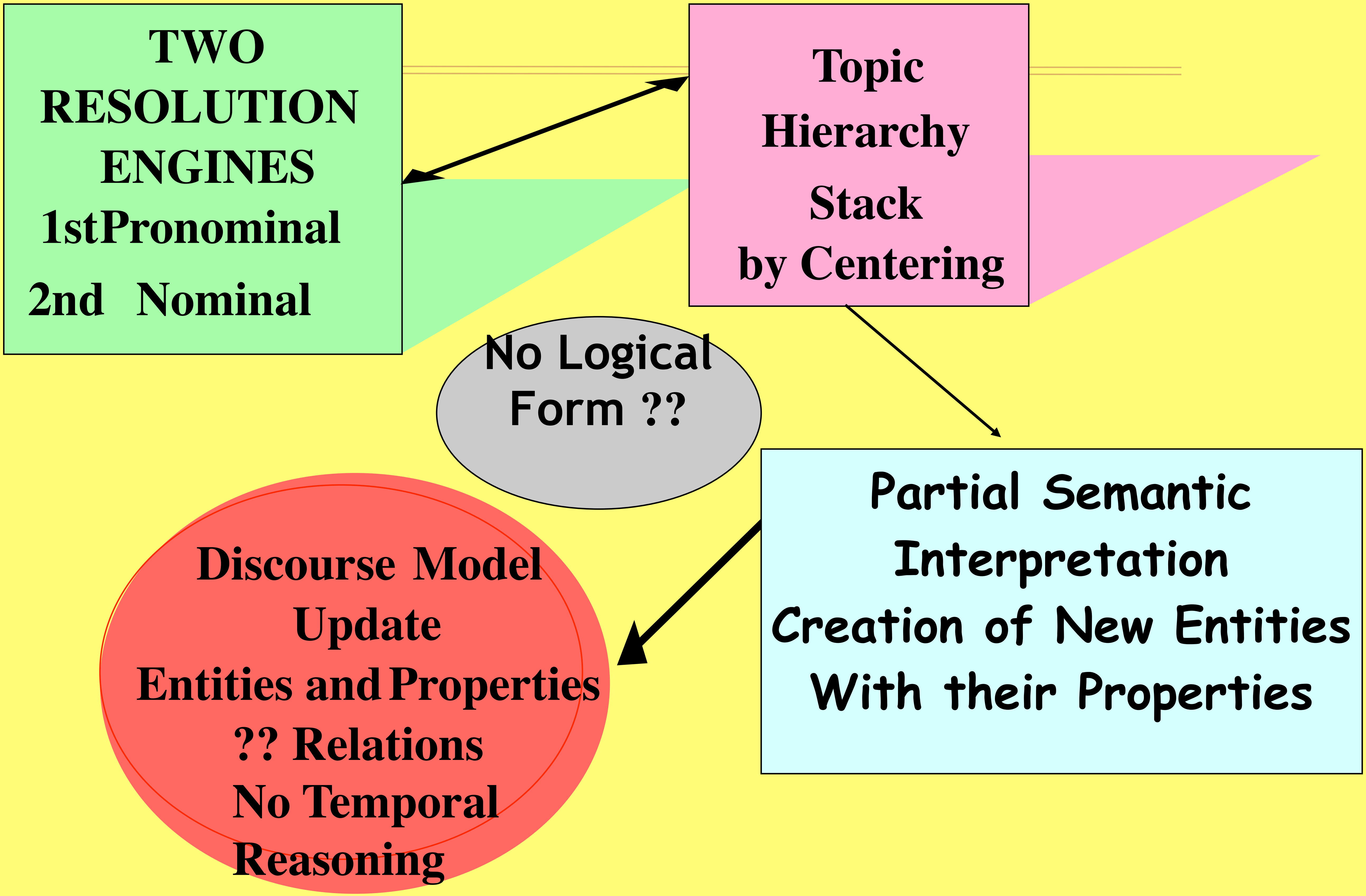
Subcategorization

- Predicate-Argument Structure

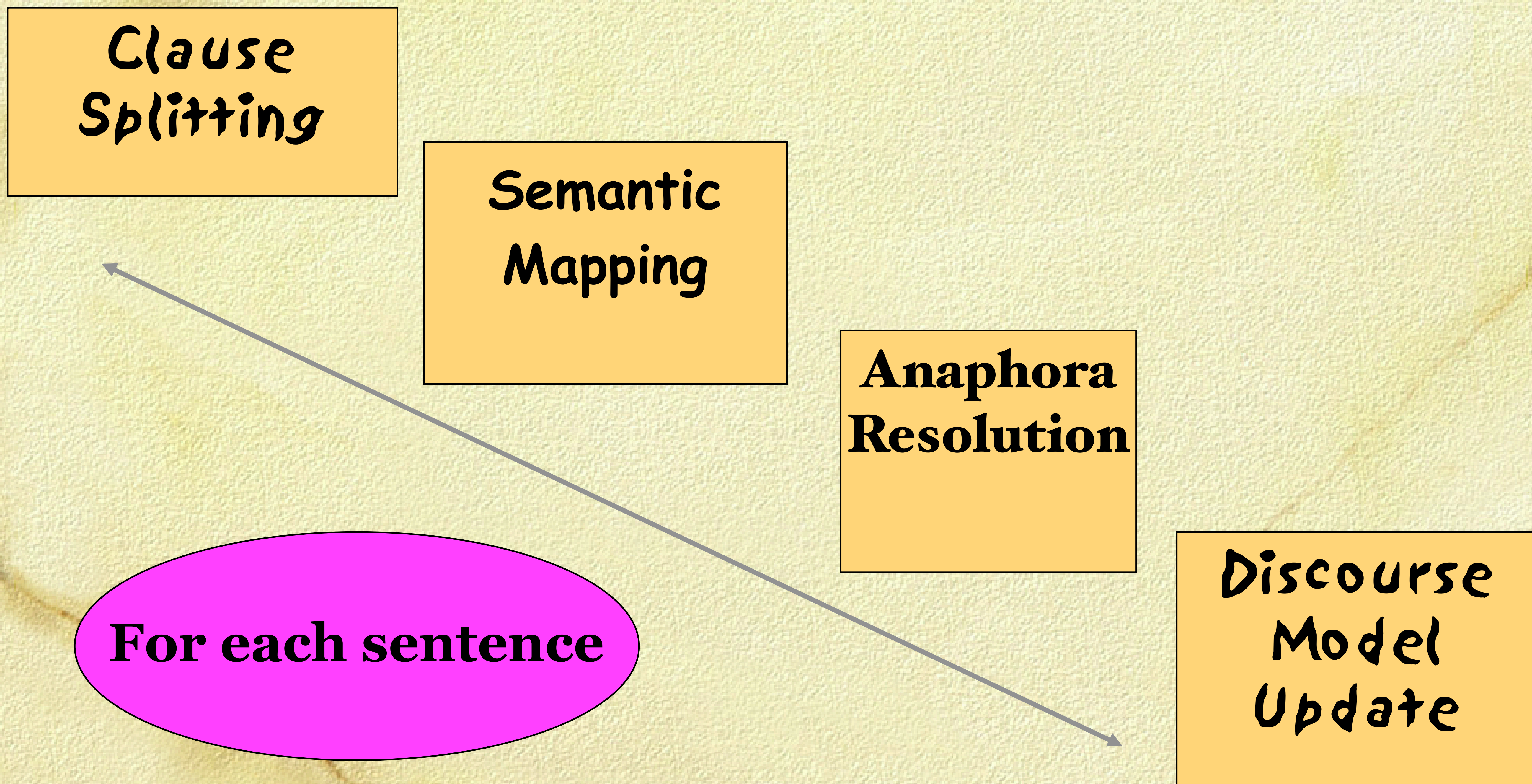
75%

- Partial Semantic Interpretation

SYSTEM ARCHITECTURE



PARTIAL SEMANTIC MAPPING



ROBUST SEMANTIC MAPPING

Clause Splitting

Semantic Mapping

For all clause

For all clauses

Anaphora Resolution

Discourse Model Update

System Pipeline

- **Repeat for each sentence**
  extract_ref_exprs(Net, RefList),
- **ref_ex(SnX/SentNo,Head,Tab,Def,Part,**
  Card,Class,Num,SCat,F/Role,Mods)
- **resolve_externals(SentNo, RefList, Args),**
  topic_hierarchy(SentNo, Args)
  end

System Pipeline

- extract_ref_exprs(Net, RefList)

  Repeat for each sentence

    collect all grammatical functions

    then, for each clause

    do,

      interpret grammatical functions

        by searching subcategorization frames associated to predicates

        associate semantic roles to arguments

            (from COMLEX)

          and semantic categories(from WordNet)

          continued...

System Pipeline

- **Continued,**
  for each clause
  associate semantic roles to modifiers
  and adjuncts also by linking to their
  governing relations (from COMLEX)
  and semantic categories (from WordNet)
  then,
  anaphora resolution
  semantic individuals and properties
  update the Discourse Model
  end

A short text from The Guardian

Thursday, 25th June 2001

National Parties and the Internet

by Joanna Crawford

A survey of how national parties used the internet as a campaigning tool during the election will brand their efforts "bleak and dispiriting" - despite the pre-campaign hype of an "e-election". Researchers from Salford University studied websites from all the major parties during the general election, as well as looking at every site put up by local candidates. Their conclusions - to be presented tomorrow at a special conference organised by the Institute for Public Policy Research - could influence how future political contests, including the forthcoming Euro debate, are carried out on the web.The report finds that none of the major three parties allowed message boards or chat rooms for users to post their opinions on the sites. It states: "Parties were accused of simply engaging in online propaganda with boring content and largely ignoring interactivity."

A short text from The Guardian

The report concludes: "The new media is a way for *them* to get closer to the public without necessarily allowing the public to become overly familiar in return. *The authors* - Rachel Gibson and Stephen Ward - go on to state that this may be because *parties* still regard the web as an electioneering tool, rather than as a democratic device. *They* said: "*Very few* offered original material, or changed *their* sites noticeably over the course of the campaign. Indeed, a large majority of *local sites* were really no more than static electronic brochures.” *They* dub this "rather disappointing", but praise the *Liberal Democrats* as "clearly the most active" with around 150 sites. The report concludes: "*Parties*, as with the general public, need *incentives* to use the technology. As yet, there seems more to lose and less to gain if *they* make mistakes experimenting with the technology."

Pronominal Expressions

- 2-their
- 4-their
- 5-none, 5-their
- 6-it
- 7-them
- 8-this
- 9-they, 9-their

10-majority

11-they, 11-this

13-they

A short text from The Guardian

Thursday, 25th June 2001

National Parties and the Internet

by Joanna Crawford

A survey of how national parties used the internet as a campaigning tool during the election will brand their efforts "bleak and dispiriting" – despite the pre-campaign hype of an "e-election".

Researchers from Salford University studied websites from all the major parties during the general election, as well as looking at every site put up by local candidates.

Their conclusions – to be presented tomorrow at a special conference organised by the Institute for Public Policy Research - could influence how future political contests, including the forthcoming Euro debate, are carried out on the web.

The report finds that none of the major three parties allowed message boards or chat rooms for users to post their opinions on the sites. It states: "Parties were accused of simply engaging in online propaganda with boring content and largely ignoring interactivity."

A short text from The Guardian

The report concludes: "The new media is a way for them to get closer to the public without necessarily allowing the public to become overly familiar in return.

The authors - Rachel Gibson and Stephen Ward - go on to state that this may be because parties still regard the web as an electioneering tool, rather than as a democratic device.

They said: "Very few offered original material, or changed their sites noticeably over the course of the campaign. Indeed, a large majority of local sites were really no more than static electronic brochures."

They dub this "rather disappointing", but praise the Liberal Democrats as "clearly the most active" with around 150 sites.

The report concludes: "Parties, as with the general public, need incentives to use the technology. As yet, there seems more to lose and less to gain if they make mistakes experimenting with the technology."

SEMANTIC INFERENTIAL NETS

- ***internet***
- **tool**
- **website**
- **site**
- **web**
- **interactivity**

sites

media

device

material

brochures

technology

CHUNKS-BASED SUMMARY

Thursday , 25/th June 2001 National_Parties and the Internet by Joanna_Crawford .

It states ':' " Parties were accused of simply engaging in online propaganda with boring content and largely ignoring interactivity .

The report concludes ':' " the new media is a way for them to get_closer to the public without necessarily allowing the public to become overly familiar in return .

The authors - Rachel_Gibson and Stephen_Ward - go_on to state that this may be because parties still regard the web as an electioneering tool , rather_than as a democratic device .

The report concludes ':' " Parties , as_with the general public , need incentives to use the technology .

PARTIAL-SEMANTICS SUMMARY

Thursday , 25/th June 2001 National_Parties and the Internet by Joanna_Crawford .

A survey of how national parties used the internet as a campaigning tool during the election will brand their efforts " bleak and dispiriting - despite the pre-campaign hype of an " e-election .

The report finds that none of the major three parties allowed message_boards or chat_rooms for users to post their opinions on the sites .

It states ':' " Parties were accused of simply engaging in online propaganda with boring content and largely ignoring interactivity .

The report concludes ':' " the new media is a way for them to get_closer to the public without necessarily allowing the public to become overly familiar in return .

COMPLETE-SEMANTICS SUMMARY

Thursday , 25/th June 2001 National_Parties and the Internet by Joanna_Crawford .

A survey of how national parties used the internet as a campaigning tool during the election will brand their efforts " bleak and dispiriting - despite the pre-campaign hype of an " e-election .

Researchers from Salford_University studied websites from all the major parties during the general_election , as_well_as looking_at every site put_up by local candidates .

Their conclusions - to be presented tomorrow at a special conference organised by the Institute for public Policy Research - could influence how future political contests , including the forthcoming Euro debate , are carried_out on the web .

Question-Answering with GETARUNS

How Maple Syrup is Made

Maple syrup comes from sugar maple trees. At one time, maple syrup was used to make sugar. This is why the tree is called a "sugar" maple tree.

Sugar maple trees make sap. Farmers collect the sap. The best time to collect sap is in February and March. The nights must be cold and the days warm.

The farmer drills a few small holes in each tree. He puts a spout in each hole. Then he hangs a bucket on the end of each spout. The bucket has a cover to keep rain and snow out. The sap drips into the bucket. About 10 gallons of sap come from each hole.

Hard to Parse Sentences

- How Maple Syrup is Made
- Maple syrup comes from sugar maple trees. At one time, maple syrup was used to make sugar. This is why the tree is called a "sugar" maple tree.
- Sugar maple trees make sap. Farmers collect the sap. The best time to collect sap is in February and March. The nights must be cold and the days warm.
- The farmer drills a few small holes in each tree. He puts a spout in each hole. Then he hangs a bucket on the end of each spout. The bucket has a cover to keep rain and snow out. The sap drips into the bucket. About 10 gallons of sap come from each hole.

DISCOURSE MODEL

- FACT is an
  Infon(Index,
    Relation(Property),
    List of Arguments - with Semantic Roles,
    Polarity - 1 affirmative, 0 negation,
    Temporal Location Index,
    Spatial Location Index)

Who collects maple sap?

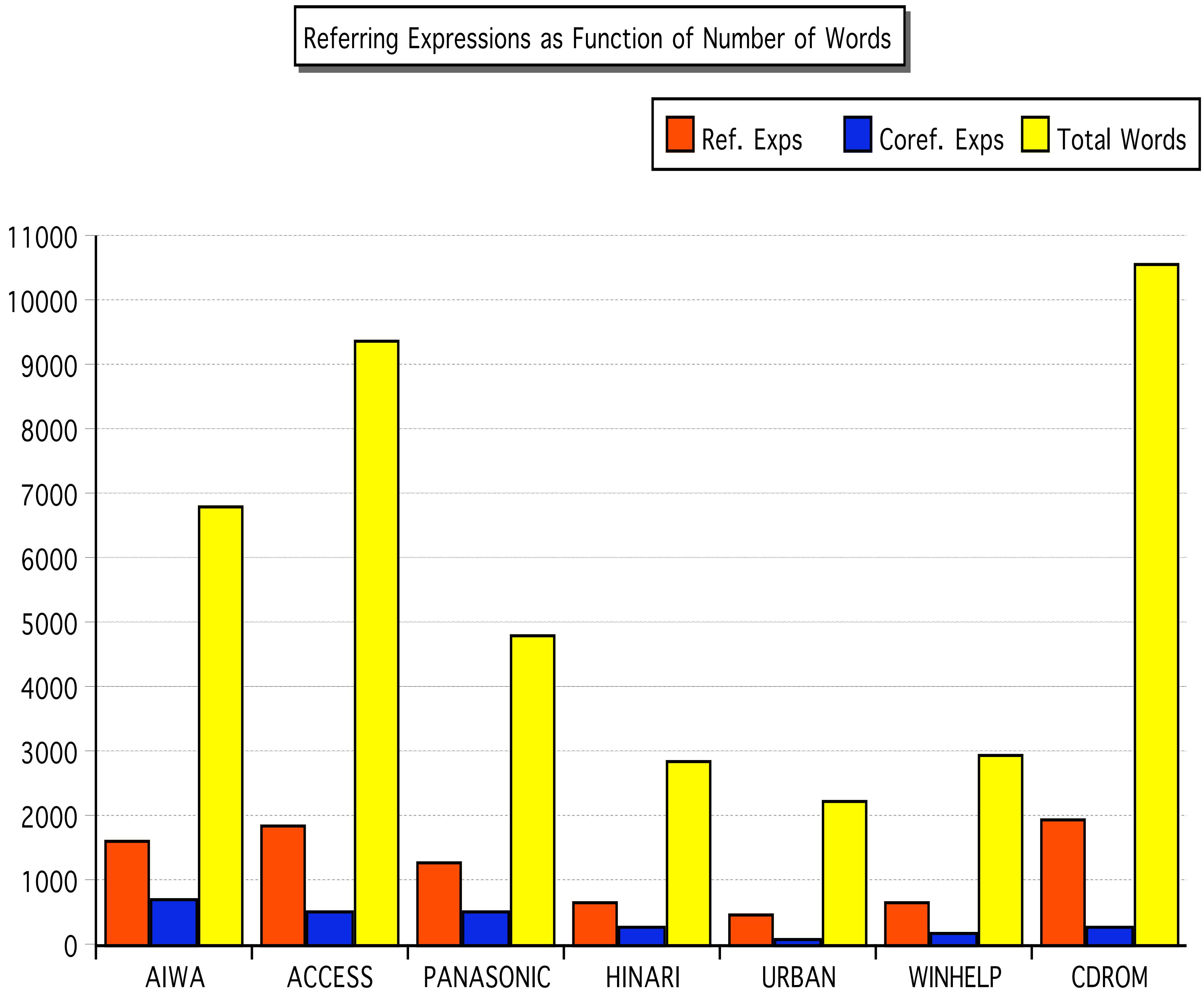
q_loc(infon3, id1, [arg:main_tloc, arg:tr(f1_uq_1)])
q_ent(infon4, id2)
q_fact(infon5, isa, [ind:id2, class:who], 1, id1, univ)
q_fact(infon6, inst_of, [ind:id2, class:man], 1, univ, univ)
q_class(infon7, id3)
q_fact(infon8, inst_of, [ind:id3, class:coll], 1, univ, univ)
q_fact(infon9, isa, [ind:id3, class:sap], 1, id1, univ)
q_fact(infon10, focus, [arg:id2], 1, id1, univ)
q_fact(infon11, maple, [ind:id3], 1, id1, univ)
q_fact(id4, collect, [agent:id2, theme_aff:id3], 1, tes(f1_uq_1), univ)
q_fact(infon13, isa, [arg:id4, arg:pr], 1, tes(f1_uq_1), univ)
q_fact(infon14, isa, [arg:id5, arg:tloc], 1, tes(f1_uq_1), univ)
q_fact(infon15, pres, [arg:id5], 1, tes(f1_uq_1), univ)

Farmers collect maple sap

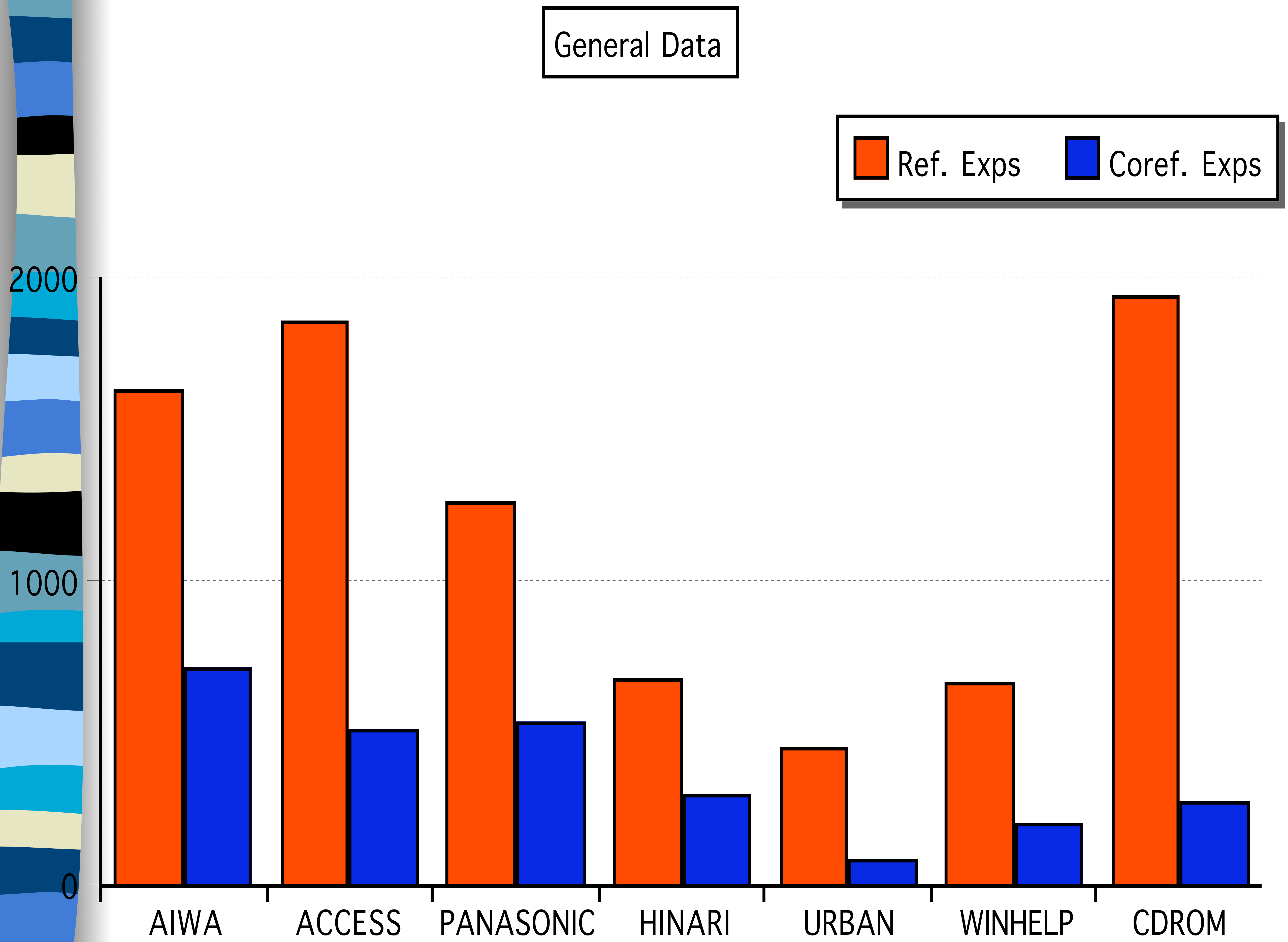
udm_loc(infon3, id1, [arg:main_tloc, arg:tr(f1_ua_1)])
udm_ent(infon4, id2)
udm_fact(infon5, isa, [ind:id2, class:farmer], 1, id1, univ)
udm_fact(infon6, inst_of, [ind:id2, class:man], 1, univ, univ)
udm_class(infon7, id3)
udm_fact(infon8, inst_of, [ind:id3, class:coll], 1, univ, univ)
udm_fact(infon9, isa, [ind:id3, class:sap], 1, id1, univ)
udm_fact(infon11, maple, [ind:id3], 1, id1, univ)
udm_fact(id4, collect, [agent:id2, theme_aff:id3], 1, tes(f1_ua_1), univ)
udm_fact(infon13, isa, [arg:id4, arg:pr], 1, tes(f1_ua_1), univ)
udm_fact(infon14, isa, [arg:id5, arg:tloc], 1, tes(f1_ua_1), univ)
udm_fact(infon15, pres, [arg:id5], 1, tes(f1_ua_1), univ)

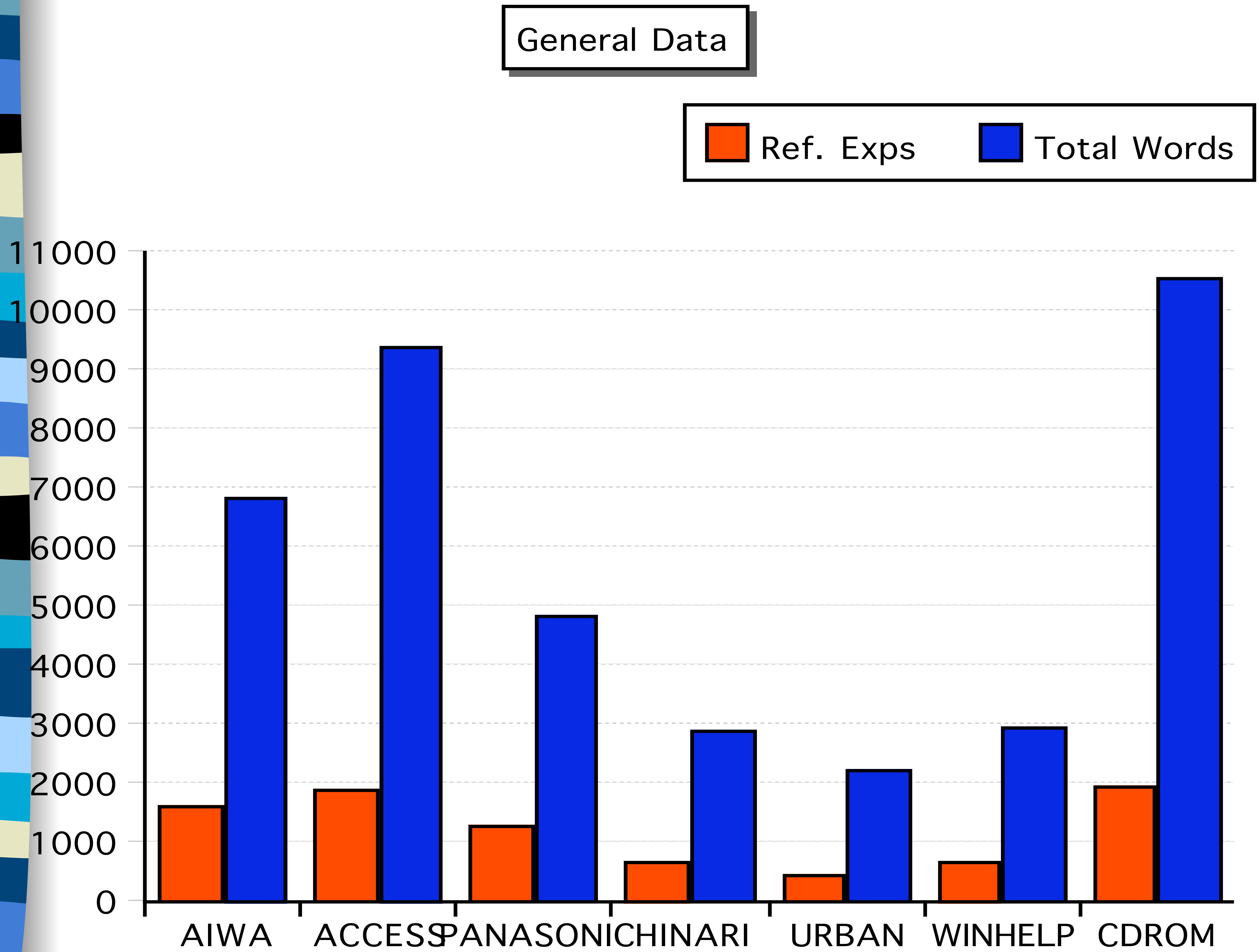
OUR PROPOSAL

- Large-scale indexing via partial parsing
- Search Engines to do IE by keywords
  - Then use top paragraph length candidates to search for answers and generate from deep analysis
  - Do the same for summaries
- Apply deep analysis to the web and produce full-fledged knowledge representation from DMs of its linguistic content

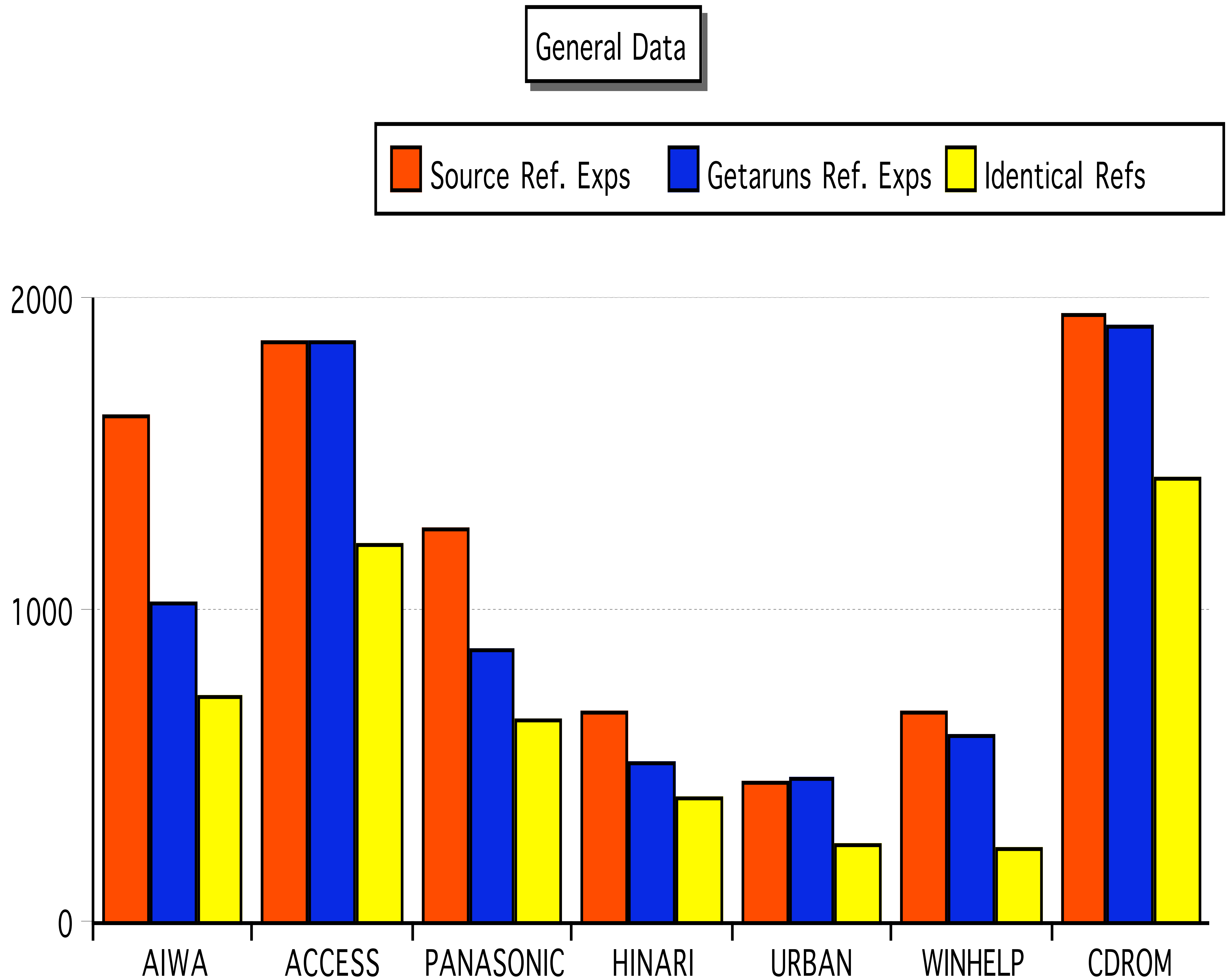


Referring Expressions as Function of Number of Words
Ref. Exps
Coref. Exps
Total Words
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
AIWA
ACCESS
PANASONIC
HINARI
URBAN
WINHELP
CDROM


General Data
Ref. Exps
Coref. Exps
2000
1000
0
AIWA
ACCESS
PANASONIC
HINARI
URBAN
WINHELP
CDROM



General Data
Ref. Exps
Total Words
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
AIWA
ACCESS
PANASONIC
HINARI
URBAN
WINHELP
CDROM


General Data
Source Ref. Exps
Getaruns Ref. Exps
Identical Refs
2000
1000
0
AIWA
ACCESS
PANASONIC
HINARI
URBAN
WINHELP
CDROM