Information Processing

Anna Maria Di Sciullo
Université du Québec à Montréal
C.P. 8888, Succursale Centre-Ville
Montréal, Québec, Canada, H2C 2P8
514-987-3000-3519#

di_sciullo.anne-marie@uqam.ca
www.interfaceasymmetry.uqam.ca

1. Purpose

The main point of this paper is that knowledge-rich information processing systems will enhance the performance of information processing, including the processing of information in the area of Digital Libraries. The term "Digital Libraries" covers the creation and distribution of all types of information over networks. People still create information that has to be organized, stored, and distributed, and they still need to find and use information that others have created. DL are a means of easily and rapidly accessing books, archives and images of various types. They are now widely recognized by commercial interests and public bodies alike. While the advantages of DL are well known [1],[2],[3],[4], the extraction of information to fill the data bases from documents, as well as the extraction of information within DL, are still a subject of research and development. While advances have been made in these areas, there are many aspects of information processing that need improvement.

2.1 Knowledge-rich properties and the form of knowledge

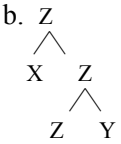
By knowledge-rich properties, we refer to the knowledge underlying the expressions of natural language, as well as the knowledge on the similarities and the differences between languages. DL now exist in a variety of languages, and the development of information processing systems incorporating knowledge-rich properties of different languages will make DL usable by different populations. The knowledge underlying text is not only rich, but it is also abstract. This knowledge is used implicitly in our understanding of texts, be they digital or not. One kind of abstract knowledge is the knowledge of the structural relations relating words and sentences, sentences in texts, and texts in collections. In fact, one of the achievements of linguistic theory, in particular in the areas of the syntax and semantics of natural language, has been to show that the parts of linguistic expressions form constituents where asymmetric relations, such as *precede* and *dominate* – identified on binary-branching trees – provide the necessary skeleton for semantic interpretation. This has been shown to be the case, word-internally, sentence-internally, as well as in the domain of text [5], [6], [7], [8]. Furthermore, the fact that sentences, texts, and collections include entities such as person and places that are, in certain cases, related to one another is another example that abstract relations underlie information structure, enabling their understanding. Moreover, the fact that languages use typical structures for asking questions about certain aspects of a knowledge domain brings additional evidence that abstract relations underlie the form of linguistic expressions for querying a knowledge domain. In languages such as English, question-words are generally sentence-initial; this is not the case in other languages such as Japanese for example. Notwithstanding the surface differences between languages, the same abstract relation supports their syntax-semantic properties. For example, at the syntax-semantic interface, a question word dominates the domain where it binds a variable, which stands for the

unknown knowledge the question asks about. The processing of abstract properties of relations is necessary for the development of technologies with a high level of accuracy.

This paper points at the necessity for information processing technologies that search unstructured texts, to fill data bases or that search within DL, to incorporate knowledge-rich properties of natural language. Rich information is necessary for information retrieval and question answering systems. This is also a necessary requirement for efficient information extraction systems such as data mining, which seek to extract entities, such as names, locations, facts, and sentiments (opinions, judgements) from texts. Most current information processing systems however do not take into account the abstract relations that are a core property of the information conveyed by the expressions in natural language. Instead, techniques that obviate these relations are used. This is the case for example for the bag-of-words technique, as well as key word, Boolean operators, and statistical calculi. We will illustrate the need to take into account the abstract relations of precedence and dominance, as well as asymmetric c(onstituent)-command, by considering the processing of argument structure, which is now starting to be recognized as central in information technology. The notion of asymmetric c-command is defined as in (1) in terms of the relations between two categories X, Y in the representations in (2), where (2a) and (2b) are equivalent.

(1) a. *C-command*: X c-commands Y iff X and Y are categories and X excludes Y, and every category that dominates X dominates Y.
b. *Asymmetric c-command*: X asymmetrically c-commands Y, if X c-commands Y and Y does not c-command X. [9]

(2) a. $[_Z$ X $[_Z$ Z Y]]

b.


Asymmetric c-command has been shown to be a core abstract relation underlying the form and the interpretation of linguistic expressions, and to govern the syntactic and semantic properties of syntactic constituents, the binding of pronouns by antecedent, as well as the properties of predicate argument structure.

2.2 Argument Structure and the form of core syntax/semantic relations

Argument structure is part of the core syntax/semantic underpinning of natural language understanding. The arguments of a predicate are the participants of the event denoted by a verbal or a nominal predicate. The core event may be modified by spatial and temporal structure external to the event, thereby restricting its denotation, or the denotation of its parts. To the difference of symbolic logic, where the arguments of a predicate are not ordered or structured with one another, in natural language, the arguments of the event are articulated on the basis of abstract relations, including precedence and dominance. That the precedence relation is relevant for information processing is illustrated in (3), where (a) and (b) do not convey the same information. That dominance relation is also relevant independently of precedence relation is illustrated in (4). In (4b) the highest nominal constituent (DP) is the external argument (subject); it is not the DP that it includes, namely *the computer scientist*. The example in (5) illustrates that the highest DP is a sentence, and not one of its sub-constituents is the antecedent of a pronominal anaphor. The example in (6) makes the same point with question-words (wh-DP), whereas (6a) is optimal, but not (6b), where the dominance relations between the wh-words are different.

(3) a. The physicist saw the computer scientist.
b. The computer scientist saw the physicist.
(4) a. [The computer scientist] met the biologist
b. [The physicist that saw [the computer scientist]] met the biologist
(5) a. [The computer scientist] introduced himself to the audience.
b. [The physicist that saw [the computer scientist]] introduced himself to the audience.
(6) a. [Who [~~who~~ saw what]]
b. [What [did [who saw ~~what~~]]]

All sentences include predicates and arguments cross-linguistically. A predicate has an argument position to saturate, and a constant (e.g., a name, a definite description) may saturate it. For example, the predicate *read* has two arguments, *the students of physics* and *The Growth of Form* in (7a). They can be questioned, as in (7b) and (7c), and pronouns can refer to them in the discourse, (7d).

(7) a. The students of physics read *The Growth of Form*.
b. What did the students of physics read?
c. Who reads *The Growth of Form*?
d. The *Growth of Form* was read several times by the students of physics. It was a real discovery for them.

Argument structure relations are asymmetric in the sense that a predicate asymmetrically selects an argument, whereas the inverse relation does not hold: an argument does not asymmetrically select a predicate. Moreover each argument is asymmetrically related to a predicate in a distinct way, in such a way that the arguments of a predicate cannot be interchanged. Thus, the predicate *read* has two arguments: the external argument of the event denoted by the predicate *read* is a DOER, and the internal argument of *read* is the internal argument (object) of the event. In the example in (7a), *The Growth of Form* is the internal argument of the event denoted by the predicate *read,* and it cannot be the DOER of the event; the external argument (subject), here the DOER of the event, is *the students*. The external argument asymmetrically c-commands the internal argument at the syntax-semantic interface – see (8), where x and y are placeholders for arguments, and pred is the placeholder for predicate.

(8) $[_{\text{Pred}}$ x $[_{\text{Pred}}$ Pred y]]

The asymmetric property of predicate argument structure holds independently of the properties of the arguments, which can be overt or null. Overt and null arguments have semantic features, while only overt arguments have physical (graphic, phonetic) features. Moreover, arguments, which we will restrict to DP (nominal arguments) for the purpose of this paper, include bare nouns, proper names, definite descriptions, indefinites, and pronouns.

Consider the examples in (9) and (10). The example in (9) illustrates the external/internal argument asymmetry with respect to the delimitation of the event denoted by a verbal predicate. The presence of a definite internal argument affects the boundedness of the event, whereas the external argument, even if it is definite, does not have such an effect. The example in (9a) denotes an activity, i.e., a rightward-unbounded event. This is evidenced by the fact that a durative adverb, such as *for one hour*, may modify the event, but not punctual adverbs such as *in an hour* – see (10). The example in (10b) illustrates the fact that a definite internal argument, *The Growth of Form*, affects the boundedness of the event denoted by the predicate by providing an

endpoint to the event. In effect, the example in (10a) illustrates that only punctual adverbs, such as *in an hour*, may modify bounded events.

(9) a. The students read.
b. The students read the *Growth of Form*.
(10) a. The students read for an hour /#in an hour.
b. The students read the *Growth of Form* for an hour /in an hour.

Precedence, dominance and asymmetric c-command relations are crucial for information processing, for intelligent information retrieval and information extraction systems, and more generally for language understanding systems.

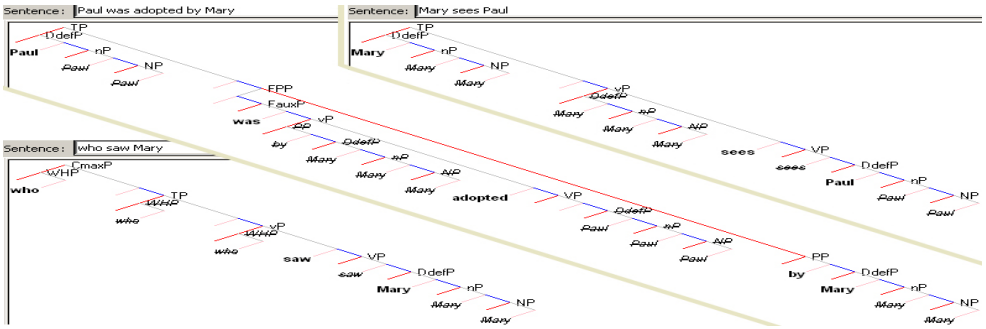
2.3 Asymmetry recovering parsers

Computational implementations of asymmetric relations are already available. The asymmetric c-command relation is part of Marcus's parser [10], as well as in the framework of Principled Based Parsing [11], [12], and in the more recent works on asymmetry and minimalism [13], [14], [15], [16]. The figure in (11) presents traces of the implemented Asymmetry Recovering Parser (ARP) [17]. One particularity of this parser is that it recovers the predicate argument structure of linguistic expression in the lower domain of the tree. It does so for any sort of linguistic expression (active, passive, question, etc.), and it is extendable to any other natural language.[1]

[1] The parser interprets the operations of the grammar as applying groups of rules in local domains. The rules of the domains (CP (Complementizer Phrase), DP (Determiner Phrase)) define the maximal realizations of syntactic projections; the rules of the groups (Grp CP, Grp DP, Grp PP (Prepositional Phrases), etc.) exhaustively enumerate the potential realizations of domains. The most inclusive domain is the CP domain, which is open by default at the beginning of the parse. Conditions, including Agreement, apply before an action is taken. After the first word has been matched with a category of the grammar, the next word's category is recovered from the numeration and is matched against the next available positions in the group. If no match is found, no parse is generated. The above procedure is repeated until a word is found that does not belong to the current group. The group is then closed, and together with it any domain it may have triggered. Following the projection path, the last domain to close is the first opened, i.e., the maximal CP domain.

Thus, a common noun is recognized as a head and is first shifted with a determiner to form a DP, before this DP is shifted in the specifier of the TP (Tense Phrase). Subsequent linkings of this DP to positions in the more inclusive domains, including the VP (Verbal Phrase) and the NP (Nominal Phrase) domains, derive the effects of movement from left to right. The parser makes an optimal use of the operations of the grammar, while it reduces the search space. The parses in (11) illustrate the output of the parser, which uniformly recovers the asymmetric relations of linguistic expressions. This result contrasts with the ones obtained by non-asymmetry recovering parsers, such as Connexor (http://www.connexor.com). Connexor does not recover the asymmetric relation between the external and the internal arguments. Furthermore, the agentive adjunct/modifier *by*-phase in the passive sentence is wrongly analyzed as an argument. Moreover in the wh-question, Connexor provides no relation between the VP internal subject position and the wh-constituent in the CP.

(11)



The recovery of asymmetrical relations by the ARP parser provides a configurational basis for the processing of the semantic relations between the constituents of linguistic expressions, including the semantic relations between the arguments of a predicate. Given the generic properties of the grammar it implements, the parser can easily handle cross-linguistic variation. Its left-to-right incremental analysis also makes it closer to human processing. Asymmetry recovering parsers can be used to enhance the performance of information processing systems, as well as the usefulness of DL.

3. Information processing

Information processing aims at designing software that will analyze, understand, and generate languages that humans use naturally. Facts from a variety of languages indicate that linguistic expressions cannot be equated with strings of characters or concatenation of words, but consist of asymmetric relations between constituents. In the following paragraphs, we discuss current practice in information processing. We point out the limits of the actual paradigm as well as the benefit of using asymmetry-recovering parsers in these systems.

3.1 Current practice

Current practice in information processing is generally based on the processing of units, such as characters and chains of characters, n-grams, without taking into account the basic asymmetric properties of natural language. This covers the whole range of applications in natural language technology from language processing systems (recognition and generation) to information content processing systems (information retrieval and extraction, question-answering systems, discourse anaphora resolution, summary production, etc.). We focus on content processing systems in order to show that current practice is not optimal, and that the processing of argument structure and modification structure asymmetries, leads to optimization.

Many natural language processing applications require the ability to recognize when two text segments, however superficially distinct, refer to specific arguments of a predicate. Information Extraction and Question Answering are examples of applications that need precise information about the relationship between different text segments with respect to predicate argument structure. This is also the case for Discourse Pronominal Anaphora Resolution and Information Retrieval. Knowledge-rich information processing systems can be used to retrieve information for and in DL, and more generally in Intranet as well as on the web more generally.

3.2 Information Retrieval

The purpose of search engines is to retrieve relevant documents based on the analysis of the queries, the analysis of a set of documents, and a method for determining the relevance of the retrieved documents [18], [19], [20].The large majority of search engines combine Boolean procedures with another method, see (13), and the retrieval of documents is based on the number of times the keywords of a query appear in the text, the keywords being related by the Boolean operators, AND, OR and NOT.

(12) a. Boolean (frequency of keywords and Boolean expression of the queries)
b. Clustering (statistical analysis grouping similar documents)
c. Linguistic analysis (stemming, synonymy-handling, spell-checking)
d. Natural language processing (named entity extraction, semantic analysis)
e. Ontology (knowledge representation)
f. Probabilistic (belief networks, inference networks, Naïve Bates)
g. Taxonomy (hierarchical relationship between concepts and categories in a particular search area)
h. Vector-based (proximity of documents and queries as arrows on a Multidimensional graph)

Furthermore, most information retrieval systems simply represent documents and queries as a 'bag-of-words'. The bag-of-words technique [21], [22], is based on the assumption that substantive words (such as *ball*, *glove*, *bat, basket*) differ fundamentally from functional words (such as *a, to, the*). In addition, it is assumed that some substantive words (e.g., *ball*) identify the relatedness of documents, while other substantive words (e.g., *glove, bat*, *basket*) identify their differences. Functional words (e.g., *a, to, the*) are irrelevant for distinguishing the documents from each other and from unrelated ones. Several problems are tied to the simple bag-of-words representation of documents, including the problems related to natural language understanding, e.g., word-sense identification, sentence and text interpretation processing. The bag-of-words approach to information processing does not take into account the asymmetry of linguistic relations. Consequently the relations between the substantive words and other substantive words mediated by the functional elements are lost, and with them the semantic content conveyed by the expressions they are a part of. In natural languages, the terms of linguistic expressions are part of asymmetric relations headed by functional heads, such as *of*, *the*, *by*, and not by Boolean operators. Thus, information retrieval systems based on keyword search and Boolean analysis will bring back wrong results for a query such as: *the funding of research by the government* including documents about '*research*' and documents about '*the funding of research'*, as well as documents that will not be relevant at all, including the terms '*government*', '*research'*, and '*funding*', each of these terms interspersed within a document. Operating search engines are not optimal. Even in the best cases, the results include irrelevant documents. The development of a new generation of search engines designed to retrieve information on the basis of the recovery of natural language asymmetries, instead of the processing of singular elements, is a step forward in the optimization of these systems.

In effect, an information retrieval system based on the recovery of asymmetric relations can handle one-word queries such as *language*, as the system may recover covert asymmetric relations. Covert asymmetric relations are independently required for the processing of null arguments of verbal predicates. An asymmetry-oriented search engine capable of recovering covert asymmetric relations will perform as well as any search engine based on keyword search. For an asymmetry recovering search engine, a one-word query is a query where a word is part of

an asymmetric relation with other covert elements. The retrieved documents include expressions where *language* is related to another term, preceding or following it. Given that *language* is a predicate, it takes an argument and it can be modified. The nominal predicate *language* is modified by *French* in (13a*)*, whereas in (13b) it takes *human* as its internal argument.

(13) a. [$_N$ French[$_N$ language y]]
b. [$_N$ human [$_N$ language [$_{Pof}$ ~~human~~]]]

Thus, one-word queries are tractable by a search engine oriented by the recovery of natural language asymmetric relations, which analyzes them minimally as predicates with covert argument structure and modification structure. Furthermore, two-word queries, such as *language detection*, can be handled by a Boolean keyword-based search engine, as well as by an asymmetry-recovering search engine. However, with the first sort of engines, documents with only one of the keywords may be retrieved, as well as documents where each one of the keywords is part of different sentences. Thus, the documents retrieved may not be directly relevant to the query.

Furthermore, a search engine that recovers asymmetric relations will analyze the query *language detection* as an asymmetric predicate argument structure relation. Only the documents about *language detection* will be retrieved. No document about one term of this relation alone, about *language* in general, or about any sort of *detection*, will be retrieved. The asymmetry between the nominal predicate *detection* and its internal argument *language* is represented in the lower layer of (14), where the copy of *language* (i.e., ~~language~~) is within the asymmetric c-command domain of the Pred$_N$ *detection*, which indicates that it is the internal argument of that predicate.

(14) [language [x detection ([$_{Pof}$ ~~language~~]]]

With a knowledge-rich search engine, the positions of arguments with respect to predicates are constant, notwithstanding their overt positions in queries and documents. In the case at hand, *language*, which is the internal argument of the predicate *detection,* precedes that predicate overtly; the asymmetry-oriented search engine assigns to the internal argument of a predicate the canonical position of the internal argument. A knowledge-rich search engine is sensitive to argument structure alternations. In the case at hand, the structure in (14) is related to the structure in (15), and the documents including expressions such as *the detection of language* will be retrieved as well as those including expressions such as *language detection*.

(15) [x [detection [$_{Pof}$ language]]]

This is not necessarily the case for operating search engines. For example, a Google search gives good results for the query *language detection*, see (16), but it gives much worse results for the equivalent query *detection of language*, see (17). Furthermore, these two queries should bring back the same results; however, different documents are retrieved.

(16) a. LanguageIdentifier.com -- Automatic **Language Detection** Software
Free software to automatically detect which languages and encodings a document is written in. Has **detection** modules for over 260 different languages and **...**
b. PHP **Language Detection** :: Detect System Languages, set headers **...**
A PHP script that will detect which languages your system has installed, then allows you to set things like headers, redirects, and cookies based on the **...**

(17) a. Robotics Institute: Large-scale Topic **Detection** and **Language** Model **...**
... K. Seymore and R. Rosenfeld, Large-scale Topic **Detection** and **Language** Model Adaptation, tech. report CMU-CS-97-152, Computer Science Department, **..**

b. ACSAC 2001 (www.acsac.org): Application Intrusion **Detection** using **...**
... Application Intrusion **Detection** using **Language** Library Calls **...** Keywords: intrusion **detection**, application, **language** library calls, signatures **...**

Finally, consider multiword queries. A Google search for a query such as *language detection by a vector machine* gives poor results, as the first two hits illustrate in (18). The retrieved documents are not directly relevant; they do not bring back documents specifically about *language detection by means of Support Vector Machines*. The first ranked, thus the most relevant, document deals with Anomaly Delection, and the second-most relevant document is a tutorial on Support Vector Machines.

(18) a. Wenjie Hu's homepage--Data Mining,Information Retrieval,Computer **...**
Robust Support **Vector** Machines for Anomaly **Detection** in Computer Security". The 2003 International Conference on **Machine** Learning and Applications **..**

b. YOV408 Programming tutorials - Support **Vector Machine** tutorials
A speech recognizer written entirely in the JavaTM programming **language** Index of motion **detection**. 24, Support **Vector Machine**,

Moreover, for an asymmetry-recovering search engine, the query in (19) has the structure in (20), where the by-phrase is a modifier related to the external argument of the predicate *detection*. The adjunct is introduced by the functional element *by*. This functional element, like *of*, does not have the symmetric properties of the Boolean operators AND and OR. In the case at hand, the functional element *by* introduces an asymmetric relation between the external argument and a modifier. This is represented by the structure in (19), where the *by*-phrase is linked to the external argument of the nominal predicate. As is the case for the examples in (20), (21) above for the identification of the internal argument, a knowledge-rich search engine will correctly identify the external argument of nominal expressions. In the case at hand, the structure in (19) and the structure in (20), with the possessive functional element *s'*, will be analyzed as equivalents. Both the possessive phrase and the *by*-phrase are linked to the external argument.

(19) [the [[x detection [of language]] [by a vector machine]]]]
(20) [a [vector machine's [detection [$_{Pof}$ language]]]]

Precision and recall benefit from a knowledge-rich search engine capable of recovering argument structure asymmetries [23], [24], [25]. This is also the case for discourse pronominal anaphora and question answering systems, as discussed in the next sections.

3.3 Pronominal anaphora resolution

Pronouns are bundles of features and their reference is dependent on an antecedent in a 'local' domain. This domain cannot be the domain of the proposition, (21), unless the antecedent of the pronoun does not asymmetrically c-command the pronoun. For example in (22), the antecedent is embedded in the subject constituent. The antecedent of a non-reflexive pronoun is generally outside of the domain of the proposition, (23). In (24), the pronoun *it* has no antecedent in the preceding proposition, as anaphora requires agreement in the morphological features of the pronouns and the antecedent.

(21) Bill appreciates him.
(22) [The books [that *Bill* reads]] make *him* appreciate information processing.
(23) John knows [that Bill appreciates him]
(24) John knows that [that Bill appreciates it]

Current discourse pronominal anaphora systems do not incorporate linguistic-rich knowledge and fail in many cases to identify pronominal reference. For example, Mitkov's Anaphora Resolution System (MARS) [26], cannot handle pronominal anaphora in several cases, as discussed in [27], [28].

Proper names and definite descriptions are possible referents to pronouns. This is also the case for indefinite expressions and quantifiers under certain conditions.

(25) Someone came to the meeting. He was expected to vote on the motion.

MARS misses the target in these cases. Based on poor linguistic knowledge, i.e., the string linear position of the constituents, and on number feature, it wrongly identifies *the meeting* as being the antecedent of the pronoun *he*:

(26) **MARS result** for (24): **He** appears in paragraph 2, sentence 2, from position 1 to position 1. It is singular. The antecedent is indicated to be **the meeting** in paragraph 2, sentence 1, from position 7 to position 8.

These results show that knowledge-poor pronominal anaphora resolution systems are not optimal. They also point to the correctness of the view that the fine-grained syntax-semantic properties of the linguistic expressions are crucial for pronominal anaphora comprehension. Knowledge-poor systems mainly process string-linear properties of linguistic expressions, and cannot handle cases where a quantifier is anaphorically related to a pronoun, as well as cases where the antecedent of the pronoun is embedded in a larger constituent.

The table in (27) shows that the performance of Discourse anaphora resolutions of knowledge-poor systems is far from optimal [29].

(27)

Researcher	Type	Success	Note
Dagan & Itai	collocations	55.93%	manually prepared corpus
Lappin & Leass (RAP)	parsed syntactic structure	72%	for intersentential anaphora
Kennedy & Boguraev	enriched POS tagging	75%	manual pre-processing
Baldwin	enriched POS tagging	77.9%	rich pre-processing
Mitkov (MARS)	shallow parsing	62.44%	fully automatic
Aone & Bennett	machine learning	76.27%	**lexical, syntactic, semantic and positional information** (rather rich)
McCarthy and Lehnert	machine learning	85.8%	manual pre-processing
Soon, Ng & Lim	machine learning	58.9%	fully automatic
Ge, Hale & Charniak	probabilistic approach	82.9%	manually tagged
Cardie & Wagstaff	clustering	52.8%	unsupervised

In Di Sciullo [30], we proposed a Pronominal Discourse Anaphora system based on the asymmetric relations relating a pronoun to its antecedent in a discourse. We summarize the results in what follows.

Assuming that a discourse is a set of conjoined propositions, and that conjunction relations are asymmetric in natural language [9], [31], [32], the propositions of a discourse are asymmetrically related, (28). Moreover, given that the propositions in a discourse, and the constituents therein, are linearly ordered, and that precedence is an asymmetric relation, if P_1

precedes P_2, the constituents in P_1, such as DPs (nominal expressions) and DPros (pronouns), precede and thus are in asymmetric relation with the constituents of P_2, where > stands for the precedence relation. Moreover, extending the Linear Correspondence Axiom, [9], to the domain of the discourse, linear precedence relations between the constituents of a discourse follows from the asymmetric c-command relations between these constituents.

(28) $[_{P1} \ \ldots DP_1 \ldots] > [\ _{P2} \ldots \ DP_2 \ldots \ ] \ > \ [_{P3} \ldots \ DP_3 \ldots .. \ ]$

We define the interface D-Linking condition on pronominal anaphora, (13), in terms of the Link operation of Asymmetry Theory, extended to apply in the Domain of the Discourse (DD). Like the other operations of this theory, Link, (14), applies under asymmetric c-command, (1), and asymmetric Agree, (31). Agree is an asymmetric relation, since proper inclusion is asymmetric. A pronoun is linked to the closest DP it asymmetrically agrees with. Pronominal anaphora resolution is essentially the identification of the closest DP/DPro with respect to which a pronoun stands in a proper inclusion relation.

(29) D-Linking (Discourse Linking)
A pronominal must be linked in its DD.

(30) *Link (α, β)*
Given two objects α and β, Link *(α, β)* creates a new object where α and β are featurally related.

(31) Agree (φ_1, φ_2)
Given two sets of features φ_1 and φ_2, Agree (φ_1, φ_2) applies if and only if φ_1 properly includes φ_2.

We take asymmetric agreement to hold in Binding,[2] since anaphors differ from their antecedents with respect to their feature structure. An antecedent properly includes the set of features of an anaphor, since without the antecedent the reference of an anaphor cannot be interpreted. The grammatical features, i.e., phi-features of anaphors and pronouns, must also be part of the linking relation along with the semantic features: no linking relation holds between the pronoun and the c-commanding DPs in (24) for example. Thus, the following question arises: What are the features that must be part of the asymmetric agreement relation in pronominal anaphora resolution?

There are three sorts of features legible at the interfaces: phonetic, formal, and semantic. The phonetic features are legible at the phonetic/visual interface, the formal and the semantic features are legible at the semantic interface and are determinant in anaphora resolution. We focus on the formal and the semantic features in what follows.

We take the elements that enter into anaphoric relations to have the formal feature D (Determiner) and the phi-features Person (Per), Number (Num), Gender (Gen), as well as morphological Case (e.g., nominative, accusative, dative, oblique), which we will not consider here. Both pronouns (DPro) and definite determiners are D, but differ in their phi-features, definite determiners not being specified for person and gender, nor for morphological case. DPs differ from DPros, Ns are inherently 3rd pers. Argument DPs and DPros have semantic features that participate in anaphoric relations. DPs have independent reference [+Ir], whereas DPros are [-Ir]. An anaphoric relation has only one [+Ir] feature, the [-Ir] feature of DPros is linked by the [+Ir] feature of the antecedent DPs. Given Binding Theory, an anaphoric pronoun such as *himself*

[2] The Binding theory [33], determines the relations between pronouns and anaphors with respect to their antecedents in the local domain of the proposition or the DP. Accord to this theory, a category A binds a category B iff A asymmetrically c-commands B and that A and B are co-indexed (assigned the same referential index), and both A and B are in argument positions. This theory includes the following two conditions: An anaphor is bound in its local domain. A pronoun is free in its local domain.

must be bound by an antecedent in its BD, whereas pronouns must be free. Given D-Linking, a pronoun such as *him* must be linked in its DD.

The necessary formal and semantic features for pronominal anaphora resolution based on asymmetric agreement are specified in (32). The feature specifications are provided for both strong and week DPros, such as the pronominal clitics of Romance languages, e.g., *le CLO le voit* (Fr.) 'the CLO sees him'. We will not discuss week pronominal anaphora here, the properties of which also follow from asymmetric agreement, (31), [7]. The semantic features include the independent reference feature ([±Ir]), along with the animate ([±ani]) feature and the part-whole ([±w]) feature. The [±ani] feature differentiates *he* from *it*, and the [±w] feature, differentiates anaphoric, such as *himself*, from non-anaphoric pronouns, such as *he* and *him*, anaphoric pronouns are [+w], non-anaphoric pronouns are [-w].

(32) DPros and DPs formal and semantic features

	Formal: pers, num, gen	Semantic: Ir, ani, w
DPro		
strong	+ + +	+/- +/- +/-
week	- + +	+/- +/- +/-
DP	3rd pers. + +	+ +/- +/-

Given asymmetric agreement, the set of features of the antecedent must be the superset of the set of features of the anaphor. This makes the correct predictions within the propositional domain, BD. It also makes correct predictions in the domain of the discourse, DD, as the examples in (33) and (34) illustrate.

(33) a. [the author] trusts [himself]]
{ +Ir, +ani, +w } { -Ir, +ani, -w}
{+3rd pers,+ sing, +mass} {+3rd prs,+ sing, +masc}

b. [the author] trusts [him]]
{+Ir, +ani, +w } {-Ir, +ani , -w}
{+3rd prs,+ sing, +masc} {+3rd prs,+ sing, +masc}

c. [[the editor] thinks [that [the author] trusts [him]]]
{ +Ir, +ani, +w} { -Ir, +ani, +w}
{3rd prs,+ sing, +masc} {+3rd prs,+ sing, +masc}

In (33a), the BD is the proposition, and the features of the pronominal anaphor *himself* are properly included in the set of features of its asymmetrically c-commanding antecedent [the author]. In (33b), the BD of the pronoun *him* is also propositional, and the pronoun is not bound by the local antecedent with which it asymmetrically agrees, by the principles of the Binding Theory. In (33c), the pronoun *him* is free in its BD, as predicted by the Binding Theory, and it must be bound outside of that domain, in its DD, given D-Linking. A possible antecedent for this pronoun is the DP in subject position *the editor* in the matrix clause. Consider now the text in (34)

(34) [[$_{LD}$[the president] talked to [the members of the company]

{+Ir, +ani, +w} {+Ir, +ani, +w}

{+3rd pers,+ sing, +masc} {3rd pers,+ plur, +masc}

today. [The reactions of the shareholders] were unequal.

{+Ir, -ani, +w }

{+3rd pers,+ plur, +masc}

[The minutes of the meeting] indicate [that [$_{BD}$ the vice-president]

{+Ir, -ani, +w} {+Ir, +ani, +w}

{+3rd pers, + plur, +neut} {+3rd pers,+ sing, +masc)

trusts [him]]]

{-Ir , +ani, +w}

{+3rd pers,+ sing, +masc}

In (34), the pronoun *him* is not bound by its local antecedent *the vice-president* with which it asymmetrically agrees in its BD, as predicted by the Binding Theory. Given D-Linking, the pronoun must be linked in its DD. The antecedent of the pronoun is the DP *the president* in the first proposition of the discourse, since it is the closest DP with which the features of the pronoun may enter into a proper inclusion relation. The intermediate DPs do not share the same set of phi-features with the pronoun. Thus, no superset relation can be established between the set of features of the intermediate DPs and the set of features of the pronoun. Thus, the facts are correctly predicted by our approach to discourse pronominal anaphora based on asymmetric agreement.

Summarizing this section, the Binding Theory is an interface legibility condition on the interpretation of anaphors and pronouns within propositions. D-Linking is a discourse interface legibility condition that requires pronouns, i.e., elements that lack independent reference, to be linked to the closest antecedent with which they asymmetrically agree. This linking relation is obtained under asymmetric agreement, as is the case for binding. Both D-Linking and Binding are determined on the basis of local domains of argument structure asymmetries. Pronominal anaphora resolution crucially relies on the identification of these domains, where asymmetric agreement holds between pronouns and antecedents. The use of knowledge-rich discourse pronominal anaphora resolution systems in the development and the exploitation of Digital Libraries is an important aspect of tracing information about a book. The chain of reference to a given book includes the use of pronouns, and their antecedents must be optimally identified across propositions. As we illustrated in this section, this is not a task that is performed optimally by knowledge-poor discourse pronominal anaphora resolution systems. In the next section, we discuss question answering systems in the same perspective.

3.4 Question answering

A question answering system takes users' requests for information and outputs some results related to the request after searching the information in some knowledge base with which the application interfaces. Historically, there have been two major types of QASs: natural language interfaces to databases (type 1), and dialogue interactive advisory systems (type 2) (see [34] for a discussion of the characteristics of each type of system).

These two types of QASs can be distinguished in terms of their knowledge bases. The knowledge bases of type-1 QASs are databases of structured information. The natural language requests that users input into type-1 systems are translated into a structured query language or SQL, which is used by the application to access the database for information. On the other hand, although the first type-2 applications used structured data as their knowledge source, they could also search for unstructured information scattered over large collections of text documents. The ability of type-2 QASs to use text documents as their knowledge source led to the design of new types of question answering applications after the advent of the World Wide Web in the early 90's, namely the design of web-based QASs or applications that can potentially use the World Wide Web as their knowledge source. For example, the START system (see [35]) uses the web only to extract information to be stored in smaller databases. The resulting smaller database then functions as the knowledge base. Some examples of current web-based QASs are Answer Bus, Ask Jeeves, IONAUT, START, QuASM, and WebQA, see [36], [37], for discussion.

Operating Question Answering systems are also linguistically knowledge-poor. Even the MIT START, which uses inverse transformations in the processing of questions, still fails to provide answers to simple forms of questions. Whereas the same answer is provided by START for the questions in (35) and (36), the system is not able to provide an answer for the question in (37).

> (35) ===> What invention is Marconi responsible for
> Marconi, Guglielmo
> *Transmitting electrical signals*—Marconi's experiments led to practical wireless telegraphy and radio. In 1901 he successfully received signals transmitted from England to Newfoundland. He was awarded the 1909 Nobel Prize for Physics. (1975)
> **Source: Infoplease.com**
>
> (36) ===> What did Marconi invent?
> Marconi, Guglielmo
> *Transmitting electrical signals*—Marconi's experiments led to practical wireless telegraphy and radio. In 1901 he successfully received signals transmitted from England to Newfoundland. He was awarded the 1909 Nobel Prize for Physics. (1975)
> **Source: Infoplease.com**
>
> (37)===> What did Marconi do?
> *Since I found too many answers to your question, I will show you just a few of them --*
> The coordinates of Marconi, California are 38.1439 N, 122.877 W.
> Marconi is a city located in the USA, in the state of California.
> The coordinates of Marconi, Hawaii are 21.7033 N, 157.977 W.
> Marconi is a city located in the USA, in the state of Hawaii.
> **Source: Infoplease.com**

These results clearly show that argument structure asymmetries are not taken into consideration by START. The success obtained for the questions in (29) and (30) is not due to the actual understanding of questions, but rather to the use of proximity calculi and pattern matching

methods. In order to process the question in (31), the system should be able to identify *Marconi* as being the external argument (agent of the event denoted by the verbal predicate), and the interrogative pronoun *what* as being the internal argument of the verbal predicate, which refers to a set of things brought about by Marconi.

An asymmetry recovering question answering system would assign the same structural relations to the constituents in the questions in (36) and (37). The structure in (38) and (39) present the argument structure asymmetries holding between the questioned internal argument interrogative pronoun *what*, originating in the predicate argument structure domain and linked to the specifier of CP, and the external argument *Marconi*, also originating in the predicate argument structure domain, where it asymmetrically c-commands the internal argument *what*.

(38) $[_{CP}$ what $[_{C}$ did $[_{TP}$ Marconi $[_{T}$ $[_{vP}$ ~~Marconi~~ $[_{VP}$ invent ~~what~~]]]]]]

(39) $[_{CP}$ what $[_{C}$ did $[_{TP}$ Marconi $[_{T}$ $[_{vP}$ ~~Marconi~~ $[_{VP}$ do ~~what~~]]]]]]

The processing of the argument structure asymmetries is an important aspect of an efficient questioning-answering system. Moreover, predicate-argument asymmetries being constant across languages, question answering systems can achieve high levels of efficiency regardless of the language of the collection it processes. Thus, in languages such as Italian, an interrogative pronoun may occupy the canonical internal argument position, e.g., *Marconi invento che?* (Marconi invented what) 'What did Marconi invent?' This is not the case in English, where an interrogative pronoun must be in the left periphery of a question. This property of languages such as Italian is directly processed by a knowledge-rich question answering system. The Italian interrogative pronoun *che* 'what' is asymmetrically c-commanded by the external argument in the predicate argument structure domain, and is linked to the covert *che* in the specifier of CP, see (40).

(40) $[_{CP}$ Che [[+wh] $[_{TP}$ Marconi $[_{T}$ invento $[_{vP}$ ~~Marconi~~ $[_{VP}$ ~~invento che~~]]]]]]

Thus, the variation in the form of questions is a function of the canonical position of the arguments in the predicate argument structure domain, in conjunction with the parameters of variation. Thus, a knowledge-rich question answering system ensures a uniform processing of question-answer pairs whatever is the language of the collection or DL.

3.3 Information Extraction

Another aspect to information processing is text mining using natural language information extraction. Information extraction systems distil structured data from natural language texts by identifying name entities and the relations between these entities. Information extraction systems can be used to extract concrete data from a set of documents, which can then be analyzed with data-mining techniques to discover more general patterns. For many applications, available electronic information is in the form of natural language documents. Thus, the problems of text mining, that is the discovery of useful knowledge from natural language text, is becoming a very important aspect of knowledge discovery. Here again, most of the knowledge that can be mined from text cannot be discovered using simple bag-of-words representations or vector-space representations.

Information extraction systems locate specific elements of data in natural language documents. Name entity recognition is one type of Information Extraction involving the identification of the reference of certain kinds of objects such as names of individuals, institutions, companies, locations, etc. Information extraction systems also aim to extract specific relations between the entities. For example, one can identify that a particular publishing house is

publishing certain specific kinds of books, and that the particular publishing house is located in a particular city, etc.

(41) publisher (book, year), book (topic, year), author (topic, book)

There are several approaches to building information extraction systems. Manually developed patterns extracting entities and relations from text rarely result in robust systems. Supervised machine learning methods have been applied to Information Extraction, [38], [39]. Another approach is to treat Information Extraction as a sequence labelling, where each word of a text is assigned a label, and to use a statistical sequence model such as a Hidden Markov Model, [40]. The model parameters are learned from a supervised training corpus and an efficient dynamic programming method can be used to determine the most probable tagging of a complete test document. Several Information Extraction systems treat text as a sequence of uninterpreted tokens, while others use natural language processing tools, such as part-of-speech taggers, [41], [42], phrase chunkers, [43], complete syntactic parser, [44], [45], [46], or knowledge bases, including lexical semantic databases such as word Net (28), which provide word classes that can be used to define more general extraction patterns, [47].

However, an automated extracted database will inevitably contain errors. One problem encountered with mining extracted data is the handling of textual variation. The use of complete syntactic parsers, where dependencies can be traced between parts of multiple-word expressions, such as proper names, for example names of authors and names of publishing houses, can reduce the error rate. Assuming that (multiple) words include asymmetric relations, as is the case for phrases and sentences, the relation between the nouns in proper names could be analyzed as a modification relation. The modifier restricts the reference of the nominal predicate, its order with respect to the predicate is variable, and it need not be overtly expressed.

(42) Turing, Alan Turing; MIT, MIT Press, The MIT Press; Cambridge, in Cambridge

A full syntactic parser also helps to reduce the error rate by using syntactic information in order to differentiate entities that are names of persons, such as authors, e.g., Turing or Alan Turing, or publishing houses, e.g., MIT Press or The MIT Press. Knowledge-rich properties of natural language, such as asymmetric relations, can also be used to differentiate different sorts of entities, for example proper names, such as Turing from institutions such as MIT, and locations such as Boston or Cambridge. In languages such as English, proper names are not generally preceded by a determiner, contrary to languages such as Italian and Spanish, where a determiner precedes the proper noun in some cases, e.g., *il professore Turing* (It.) vs. *Professor Turing*. This contrasts with publishing houses, which may be used without the presence of a determiner. These two cases yet are different from locations, which may or may not be introduced by a preposition. Here again we see that the internal asymmetric relations shaping entities are part of the rich knowledge that underlies expressions in natural language. This knowledge is likely to improve the performance of information extraction and data mining systems. Likewise, the use of asymmetry for the identification of relations between entities, such as the ones in (42), can be improved by taking into consideration the syntax-semantic properties of the nominal predicates, and using templates that include predicate-argument structures as well as modification relations. These relations are constant across languages, while all the variables they contain might not be spelled out overtly in texts. The recovery of asymmetric syntax-semantic properties of natural language expressed in texts helps to sharpen information retrieval and data mining systems, and brings human knowledge of languages and of the differences between languages to use in information processing.

4. Conclusions

In this paper, we have discussed an approach to using rich natural language knowledge in information systems performing information retrieval, pronominal anaphora resolution, question answering, and information extraction for text mining. These systems are relevant for search, extraction and mining of text in general, as well as being relevant for DL. The use of abstract natural language knowledge and in particular asymmetric syntax-semantic relations comes to be seen as an important dimension of efficient information processing. We have provided evidence that meaningful information is articulated on the basis of asymmetric syntactic-semantic relations. Research in information processing incorporating these abstract relations will contribute to upgrading systems for the identification of entities and relations in text, including DL. We provided empirical evidence that asymmetric relations, couched in terms of precedence, dominance, and asymmetric c-command relations between predicates, arguments and modifiers, must be recovered in order to determine the set of documents satisfying the referent of a query, to identify possible antecedents of pronouns, to provide relevant answers to questions, and to extract relevant entities and relations between entities. Information processing systems oriented by the recovery of asymmetric relations contributes to the development of efficient systems, since it relies on universal properties of natural language. Understanding natural language by humans is based on the recursive processing of asymmetric relations ensuring the mapping between form and interpretation. Information technologies recovering these relations help to bring natural language processing closer to human performance.

5. Acknowledgements

This work is supported in part by a grant from the SSHRC of Canada to the MCRI on Interface Asymmetries 214-2003-1003, www.interfaceasymmetry.uqam.ca, and by a grant from FQRSC to the Dynamic Interface project, both awarded to Professor Anna Maria Di Sciullo at the Université du Québec à Montréal.

6. References

[1] D.I.Greenstein and S.E. Thorin. *The Digital Library: A Biography*. Digital Library Federation, 2002.

[2] L. Candela et al.. *The DELOS Digital Library Reference Model – Foundations for Digital Libraries*, 2008.

[3] R.E. Kahn and V.G. Cerf. *The Digital Library Project Volume I: The World of Knowbots: An Open Architecture For a Digital Library System and a Plan For its Development*. Reston, VA: Corporation for National Research Initiatives, 1988.

[4] E.A. Fox. *The Digital Libraries Initiative – Update and Discussion*, Bulletin of the American Society of Information Science, Vol. 26, No 1, October/November, 1999.

[5] N. Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

[6] N. Chomsky. *The Minimalist Program*. MIT Press, Cambridge, MA, 1995.

[7] A.M. Di Sciullo. *Asymmetry in Morphology*. MIT Press, Cambridge, MA, 2005.

[8] D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, 2000.

[9] R. Kayne. *The Antisymmetry of Syntax*. MIT Press, Cambridge, MA, 1994.

[10] M. Marcus. *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, MA, 1980.

[11] R. Berwick, S. Abney, and C. Tenny (eds.). Principle-based parsing: computation and psycholinguistics. *Studies in Linguistics and Philosophy*. Kluwer, Dordrecht, 1991, 1-37.
[12] S. Fong. *Computational Properties of Principle-Based Grammatical Theories*. Ph.D. Thesis, Artificial Intelligence Laboratory, MIT, 1991.
[13] A.M. Di Sciullo. Parsing asymmetries. *Natural Language Processing.* Springler Computer Science Press, 2000, 24-39.
[14] A.M. Di Sciullo and S. Fong. Morpho-syntax parsing. In A.M. Di Sciullo (ed.). *UG and External Systems: Language, Brain and Computation*. 247-268. John Benjamins, Amsterdam-Philadelphia, 2005.
[15] S. Fong. Computation with probes and goals. In A.M. Di Sciullo (ed.). *UG and External Systems: Language, Brain and Computation*. 311-334. John Benjamins, Amsterdam-Philadelphia, 2005.
[16] H. Harkema. Minimalist languages and the correct prefix property. In A. M. Di Sciullo (ed.). *UG and External Systems. Language, Brain and Computation*. 289-310. John Benjamins, Amsterdam-Philadelphia, 2005.
[17] A.M. Di Sciullo, P. Gabrini, C. Batori, and S. Somesfalean, in press. Asymmetry, the grammar, and the parser. *SLI,* 2010.
[18] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
[19] T. Strzalkowski (ed.). *Natural Language Information Retrieval*. Kluwer, Dordrecht, 1999.
[20] W.B. Frakes and R. Baeza-Yates. *Information Retrieval*. Prentice Hall, 1992.
[21] D. Lewis. Naïve (Bayes) at forty: the independence assumption in information retrieval. *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Chemnitz, DE: Springer Verlag, Heidelberg, DE. 4-15. 1998.
[22] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.
[23] A.M. Di Sciullo. Understanding natural language. *Proceedings of the Fourth International Conference on Intelligent Computing and Information Systems*. Ain Shams University. Cairo, Egypt, 2009.
[24] A.M. Di Sciullo. Natural Language Asymmetry in Internet Infrastructure. *International Journal of Electronic Business* 3. Special Issue on: Multidisciplinary, Interdisciplinary and Transdisciplinary Research in Electronic Business. 328-238. 2005.
[25] A.M. Di Sciullo. Lexical Semantic Theory Based on Natural Language Asymmetry and Consequences for Information Retrieval. *WSSL*, University of Pisa, 2001.
[26] R. Mitkov. *Anaphora Resolution*. Pearson Education, Edinburgh-London., 2002.
[27] A.M. Di Sciullo. Handling pronouns intelligently. In H. Fujita and M. Mejri, (eds.), *New Trends in Software Methodologies, Tools and Techniques. Frontiers in Artificial Intelligence and Applications*. Vol. 129: 207-225. Oxford: IOS Press. 2005.
[28] Processing bound pronouns. *Enformatika.* Prague, Czech Republic. 2006.
[29] C. Batori. Optimisation du traitement automatique des relations anaphoriques par la modélisation des asymétries d'interface; le cas de l'anaphore pronominale discursive. Doctoral exam. Université du Québec a Montréal, 2009.
[30] A.M. Di Sciullo. Domains of argument structure asymmetries. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 316-320. Orlando, Florida, 2005.
[31] A. Munn. *Topics in the Syntax and Semantics of Coordinate Structures*. Doctoral dissertation, University of Maryland, 1993.
[32] A. Moro. *Dynamic Antisymmetry*. MIT Press, Cambridge, MA. 2000.
[33] N. Chomsky. *Lectures on Government and Binding*. Dordrecht.

[34] B. Green, A. Wolf, N. Chomsky, and K. Daugherty. BASEBALL: an automatic question answerer. *Proceedings of the Western Joint Computer Conference*, 219-224. Reprinted in Grosz et al. (eds.). *Readings in Natural Language Processing*. 1961.

[35] B. Katz, G. Borchardt, and S. Felshin. Natural language annotations for question answering**.** *Proceedings of the 19th International FLAIRS Conference (FLAIRS 2006)*, Melbourne Beach, FL, May, 2006.

[36] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, A.J. McFarland, and B. Temelkuran. Omnibase: Uniform access to heterogeneous data for Question Answering. *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*, June, 2002.

[37] B. Katz. Annotating the World Wide Web using natural language**.** *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*, 1997.

[38] M.E. Calif and R.J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *Journal of Machine Learning Research*, 4:177-210, 2003.

[39] R.J. Mooney and L. Roy. Content-based book recommendation using learning for text categorization. *Proceedings of the Fifth ACM Conference in Digital Libraries*, 195-204. San Antonio TX, June, 2000.

[40] L.R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE, 77(2)*, 257-286, 1989.

[41] K.W. Church. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the Second Conference on Applied Natural Language processing*, 136-143, Association for Computational Linguistics, Austin, TX, 1988.

[42] E. Brill. Transformational-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21(4), 543-565, 1995.

[43] T.A. Ramshaw and M.P. Marcus. Text chunking using transformational-based learning. *Proceedings of the Third Workshop on Very Large Corpora*, 1995.

[44] M.J. Collins. Three generative, lexicalized models for statistical parsing. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (ACL-97), 16-23, 1997.

[45] A. Culotta and J. Sorensen. Dependency tree kernels for relation extraction. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, July, 2004.

[46] S. Ray and M. Craven. Representing sentence structure in hidden Markov models for information extraction. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence. (IJCAI-2001)*, 1273-1279, Portland, OR, 1996.

[47] C.D. Fellbaum. *WordNet. An Electronic Lexical Data-base*. MIT Press, Cambridge, MA, 1998.