A Reason to Optimize Information Processing with a Core Property of Natural Language

Anna Maria Di Sciullo
Université du Québec à Montréal/Department of Linguistics, Montreal, Canada
Di_sciullo.anne-marie@uqam.ca

***Abstract*—We focus on a property of natural language enabling the processing of information conveyed by linguistic expressions: structural asymmetry. We provide evidence that structural asymmetry is a property of argument structure. We focus on Information Retrieval and Question Answering systems and we provide evidence that these systems fail to recover natural language argument structure asymmetrical relations and thus they may fail to retrieve relevant documents from large databases and to provide relevant answers to questions. The processing of the underlying asymmetric relations will contribute to the optimization of Information Retrieval and Question Answering systems.**

I. Introduction

The goal of Natural Language Processing (NLP) techniques is to build software that can efficiently process linguistic expressions. Some NLP techniques process linguistic expressions in terms of strings of words without taking into consideration the syntax-semantic properties of the structure in which they are part. Such techniques fail to process a core feature of linguistic expressions, namely the fact that linguistic expressions are formed of related constituents in asymmetrical relations.

Linguistic expressions cannot be analysed in terms of strings of words. We illustrate this with the following two examples. First, such analyses does not account for the relation between a name and a definite description, which may refer to the same individual. This is the case, for example, of the name *Wittgenstein* and the definite description *the author of the Tractatus Logico Philosophicus* refer to the same individual. Second, the relation of a pronoun to its antecedent cannot be based on linear precedence, because the relations between the parts of a linguistic expression play a role in anaphora resolution. For example, in the expression *this student of Ludwig's thinks that he is intelligent*, the pronoun *he* cannot take *Ludwig* as its antecedent, even though it is the closest possible string-linear antecedent. The antecedent of the pronoun *he* may only be the whole nominal constituent: *this student of Ludwig's*. These examples illustrate that the form and interpretation of linguistic expressions is based on properties of relations rather than on properties of strings.

Properties of relations are central in several language-related areas, including theoretical and computational linguistics. Strong hypotheses on the asymmetric (irreversible) properties of linguistic relations are central in grammar (Chomsky [1], [2], Di Sciullo and Williams [3], Kayne [4], [5], Moro [6], [7], Hale and Keyser [8], van der Hulst and Ritter [9], [10], Raimy [11], [12], Roeper [28]). The recognition of the core role of asymmetric relations in grammar has led to the elaboration of a model where the primitives are minimal asymmetric relations (Di Sciullo [13], [14], [15]). NLP processing oriented by the recovery of asymmetric relations may lead to the development of efficient software attuned which the properties of human cognitive processing.

In this paper, we focus on the recovery of the argument structure, i.e., the relations between the arguments of a predicate, and we show that natural language technologies must access this information in order to be efficient. Given that the relations between the arguments of a predicate are asymmetric, an efficient NLP system should be oriented by the recovery of argument structure asymmetries.

The organization of this paper is the following. First, we define the notion of asymmetry. Second, we illustrate argument structure asymmetry. Third, we show that the recovery of argument structure asymmetry in Information Retrieval and Question Answering is crucial in efficient information processing. Finally, a broader consequence is drawn for information processing.

II. Asymmetry

Asymmetry is a property of relations in a set such that there are no ordered pairs in that set whose members are inverted, see (1). Linguistic expressions can be represented in terms of oriented graphs, where asymmetric relations are defined in terms of 'precede', 'dominate', and 'asymmetric c-command', see (1). These properties are core properties of linguistic relations at play across the board in grammar, as well as in argument structure identification, binding, and agreement relations. For example, in the tree in (2), X asymmetrically c-commands Y.

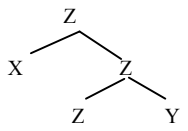
(1) a. If $R \subseteq A\ X\ A$, then R is symmetric iff
$(\forall x\ y)\ (<x, y> \in R \rightarrow <y, x> \in R)$.

b. If $R \subseteq A\ X\ A$, then R is asymmetric iff
$(\forall x\ y)\ (<x, y> \in R \rightarrow <y, x> \notin R)$. (Wall [16])

c. *C-command*: X c-commands Y iff X and Y are categories and X excludes Y, and every category that dominates X dominates Y. (Kayne [4])

d. *Asymmetric c-command*: X asymmetrically c-commands Y, if X c-commands Y and Y does not c-command X. (Kayne [4])

(2)



Asymmetry Theory (Di Sciullo [13], [17]) accounts for the fact that a change in asymmetric relations gives rise to either gibberish or a difference in semantic interpretation. We provide to examples to justify this claim. First, morphological relations are prototypically asymmetrical. Generally, the parts of a morphological object cannot be inverted without giving rise to gibberish e.g. *proto-tipic-al* vs. **al-tipic-proto*, **tipic-al-proto*, **al-proto-tipic*, **tipic-proto-al*. In cases where inversion is possible, there is a difference in semantic interpretation, e.g., from Italian *tavolinetto* (tavolo-ino-etto) 'small table' vs. *tavolettina* (tavola-etta-ina) 'small piece of wood'. Second, syntactic relations are also asymmetric: when the inversion of the constituents does not yield gibberish, semantic relations and information structure are altered. For example, in the expression *everybody loves somebody* a pair-set reading is available, such that everyone loves a potentially different person, however; in *somebody loves everybody*, a pair-set reading is not available. The change in asymmetric relations brings about a change in information structure, as the following example illustrates, *Ulysses loves Eurydice* vs. *It is Eurydice that Ulysses loves*. See Di Sciullo [13] for further discussion.

Asymmetry Theory includes a set of operations that derives interface representations that must be interpretable/legible by the performance systems. Given the central role of asymmetry in the interpretability of linguistic expressions, the following legibility condition must hold at the interfaces of the grammar and the performance systems.

(3) *Interface Legibility Condition*
Only asymmetric relations are optimally interpretable by the performance systems.

Asymmetry Theory has implications for natural language technologies, including information retrieval and question answering. While knowledge-engineered and statistical/machine-learning techniques are used to disambiguate and respond to natural language input, we expect that the processing of the asymmetric properties of linguistic expressions to enable any area where human users can benefit by communicating with their computers in a natural way.

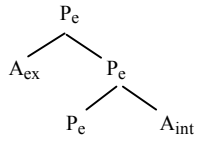
Argument structure is interpretable at the syntax-semantic interface. The arguments of a predicate, mostly nominal constituents, can be thought of as being the participants of the event denoted by the verbal or the nominal predicate. The arguments of the event are asymmetrically related, since it is not possible to interchange them without affecting the interpretation of the event. Thus *David killed Goliath* denotes a different event than *Goliath killed David*. This property of arguments is expected, if we assume that asymmetry is a basic property of linguistic expressions.

All sentences include predicates and arguments. A predicate has argument positions to saturate and a constant (e.g., a name, a definite description) may saturate an argument position of a predicate. For example, the predicate *read* has two arguments, *the students of physics* and *the Tractatus* in (4a). They can be questioned, as in (4b) and (4c), and pronouns can anaphorically be related to them in the discourse, (4d).

(4) a. The students of physics read the Tractatus.
b. What did the students of physics read?
c. Who reads the Tractatus?
d. The students of physics read the Tractatus. 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, such 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 object of the event. In the example in (4a) *The Tractatus* is the internal argument of the event denoted by the predicate and it cannot be the doer of the event; the external argument of the event is *the students*. The external argument asymmetrically c-commands the internal argument of a dyadic predicate, see (5).

(5)



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 phonetic features. Moreover, arguments, which we will restrict to DP (nominal arguments), include bare nouns, proper names, definite descriptions, indefinites, and pronouns.

Predicate argument structures denote events. Consider the examples in (6) and (7). The example in (6) 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 (6a) 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, (7a). The example in (6b) illustrates the fact that a definite internal argument, *the Tractatus*, affects the boundedness of the event denoted by the predicate by providing an endpoint to the event. In effect, the example

in (7b) illustrates that only punctual adverbs, such as *in an hour*, may modify bounded events.

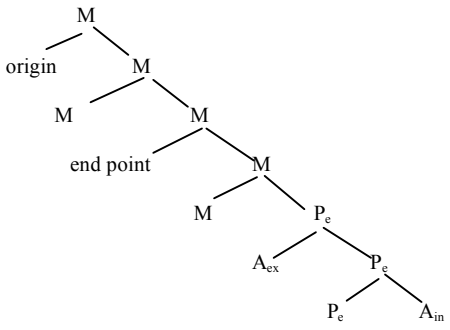
(6) a. The students read.
b. The students read the Tractatus.
(7) a.The students read for an hour /#in an hour.
b.The students read the Tractatus #for an hour /in an hour.

The core event may be modified by spatial and temporal adjunct structure external to the event denoted by the verbal predicate and its argument. Asymmetry is also a property of the spatial modification of the event. This can be illustrated with locative modifiers of the event pointing to the origin or the endpoint of the event, as the examples in (8) illustrate. Asymmetry Theory correctly predicts that there should be an asymmetry between the origin and the endpoint of an event. This prediction is borne out, as the examples in (9) illustrate, using the difference between durative and punctual adverbs. A modifier denoting a spatial origin of an event does not delimit the event; a modifier denoting the spatial endpoint of the event does have an effect on the boundedness of the event.

(8) a. The students walked from the lake to the river.
b. The students walked from the lake.
c. The students walked to the river
(9) a. The students walked from the lake for an hour/#in an hour.
b. The students walked to the river #for an hour/in an hour.

Assuming that modification relations command the basic predicate argument structure relations, we have the following configurational asymmetry between a modifier denoting the spatial origin of an event and a modifier denoting the spatial endpoint of an event. The configuration in (10) expresses the configurational locality of the spatial endpoint modifier to the basic predicate argument structure, and ensures that this sort of modifier, as opposed to the modifier denoting the origin of the event, may have an effect on the boundedness of the event.

(10)



Computational implementations of asymmetric relations are available. The asymmetric c-command relation is part of Marcus's [18] parser, as well as in Government and Binding implementations (Berwick and Weinberg [19], Berwick [20], Berwick, Abney, and Tenny [21], Fong [22]), and in the more recent works on asymmetry and minimalism (Di Sciullo [23], Di Sciullo and Fong [24], Fong [25], Harkema [26]). A computational model based on the recovery of asymmetric relations may lead to a new paradigm in natural language technology as well as to more efficient information technologies.

III. CURRENT PRACTICE IN INFORMATION PROCESSING

The current practice in information processing is based on the processing of units, such as characters and chains of characters, as well as words without taking into account the basic asymmetric properties of natural language expressions, thus leading to non-optimal results. 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, summary production, etc.). We focus on content processing systems in order to illustrate that the processing of natural language asymmetries and in particular argument structure asymmetries may lead to the optimization of these systems.

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. We focus on these two applications, even though our point covers any NLP application.

A. Information Retrieval and Extraction Systems

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 with respect to the queries (Baeza-Yates and Ribeiro-Neto [27], Strzalkowski [28], Frakes and Baeza-Yates [29]). The large majority of search engines combine Boolean procedures with another method, see (11), 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.

(11) 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)

Operating search engines give however poor results. Even the best performances include irrelevant documents. This situation can be attributed to the predominantly rich probabilistic and the poor linguistic knowledge method used by operating search engines. The development of a new generation of search engines designed to retrieve information on the basis of the recovery of asymmetric relations, instead of the processing of singular elements, is a step forward in the optimization of these systems.

A putative problem for an Information Retrieval and Extraction system based on the recovery of natural language asymmetric relations is one-word queries. A relation requires minimally two terms and a one-word query apparently does not meet this requirement.

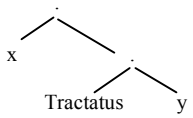
Operating search engines based on keyword search and Boolean relations retrieve a large amount of documents on the basis of a query such as (12). A Google search brings back 263,000 results, most documents on Wittgenstein's *Tractatus Logico Philisophicus*, as well as other documents addresses, including addresses for the most ancient manuscript containing the text of prayers, the *Tractatus Sacerdotalis* (treatise for the priests, Lat. 1503), and for Schlieper's *Tractatus Satanicus*.

(12) Tractatus

An Information Retrieval and Extraction system based on the recovery of asymmetric relations can handle one word, unrestricted, and thus vague queries such as (12), as it may recover covert asymmetric relations. Covert asymmetric relations are independently required for the processing of null arguments, as seen in the previous section. An asymmetry-oriented search engine capable of recovering covert asymmetric relations will perform as well as any search engine based on keyword search, since words are part of documents, documents consist of sequences of sentences, and sentences consist of constituents, whose terminal elements, i.e., the actual words of the sentences, are part of asymmetric relations.

Thus, given Asymmetry Theory, a one-word query is a query where a word is part of an asymmetric relation with another unspecified (covert) element. Thus, the retrieved documents include expressions where *Tractatus* is related to another unspecified term, preceding or following it, the term of the query is asymmetrically related to that term, as partially illustrated in (13) with constituent structure, and in (14) with ordered pairs. The asymmetry is formal and semantic, since *Tractatus* is part of an argument structure or a modifier relation. In (15a), it is the internal argument of the predicate *write*, and in (15b) it is part of the modifier of the event denoted by the predicate.

(13)



(14) <x, Tractatus>, <Tractatus, y>

(15) a. Wittgenstein wrote the Tractatus.
b. With the Tractatus in her hand, Mary went to the meeting.

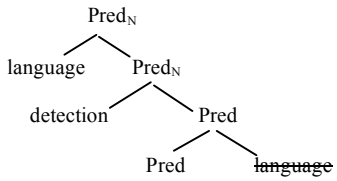
Thus, one-word queries are then not problematic for an Information Retrieval and Extraction system based on the recovery of asymmetric relations.

Two-word queries, such as (16), are handled by Boolean keyword-based search engines, as well as by asymmetry-oriented engines. However, with the first sort of engines, documents with only one of the keywords will be retrieved, as well as documents where each one of the keywords is part of different sentences. Boolean operators such as AND and OR are symmetric, consequently the keywords may appear in inverse order in the same sentence. Thus, the documents retrieved may not be directly relevant to the query.

(16) language detection

A search engine that recovers asymmetric relations will analyze the query in (16) as an asymmetric predicate-argument relation. Given that predicate-argument relations are asymmetric, 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 (17), where the copy of *language* (i.e., ~~language~~) left by the displacement of this category to a superior position within the asymmetric c-command domain of the $Pred_N$ indicates that it is the internal argument of the predicate *detection*.

(17)

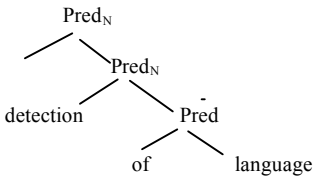


In a search engine based on fine-grained universal linguistic properties, such as argument structure asymmetry, the position of arguments with respect to predicates is constant, notwithstanding their overt position in the documents. In the case at hand, the internal argument of the predicate detection precedes the predicate overtly; the asymmetry-oriented search engine assigns to the internal argument of a predicate the canonical position of internal argument, see (5) above.

The asymmetry-oriented search engine covers the cases that can be handled by keyword Boolean search by requiring that the words be related by a natural language

operator, such as the operator OF. The structure in (17) is related to the structure in (18). Consequently, in the case at hand the documents where expressions such as *the detection of language(s)* occur will also be retrieved, among the ones where *language detection* occurs.

(18)



A Google search gives good results for the query *language detection*, see (19), but it gives much worst results for the equivalent query *detection of language*, see (20).

(19)

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 **..**

c. ASPN : Python Cookbook : **Language detection** using character trigrams
Active State Open Source Programming tools for Perl Python XML xslt scripting with free trials. Quality development tools for programmers systems **...**

d. [thelist] **Language Detection** vrs. Splash Screen
... it is to rely on **detection** of the **language** setting of the user's browser. **...**
Does anyone know why this **language detection** is not utilized more often? **...**

(20)

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 **...**

c. Information Filtering, Novelty **Detection**, and Named-Page Finding **...**
... 2.0: Information Filtering, Novelty **Detection**, and.. - **Language**.. (Correct)
0.5: **Language** Models and Structured Document Retrieval - Paul Ogilvie Jamie **...**

d. Information Filtering, Novelty **Detection**, and Named-Page Finding **...**
... 2.2: Information Filtering, Novelty **Detection**, and.. - **Language**.. (Correct)
0.3: Parsimonious **Language** Models for Information Retrieval - Hiemstra, **...**

The referent of a many-word query become harder to process by operating search engines because they are based on poor linguistic knowledge. The operating systems use shallow linguistic analysis, which is generally limited to spell-checking, stemming, part-of-speech tagging, and NP detection. Natural language processing is often reduced to named entity extraction, and the semantic analysis is limited to handling synonymy. Consider the query in (21).

(21) language detection by a vector machine

A Google search for a query such as the one in (21) gives poor results, as the first 5 hits illustrate. The retrieved documents are not directly relevant; they do not bring back documents specifically about language detection using Support vector machines.

(22)

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 **.**

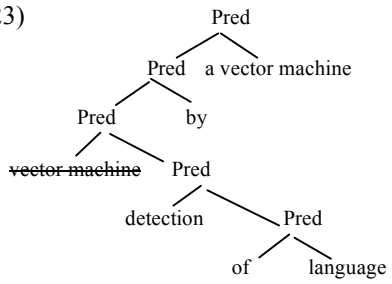
b. YOV408 Programming tutorials - Support **Vector Machine** tutorials
A speech recognizer written entirely in the JavaTM programming **language ...**
Index of motion **detection**. 24, Support **Vector Machine**, **...**

c.A discrete Kernel Approach to Support **Vector Machine** Learning in **...**
approach to Support **Vector Machine** (SVM). learning to do **Language** Independent Named **...** Protein homology **detection** using. string alignment kernels. **...**

d. A **Machine** Learning Based Approach for Table **Detection** on The Web
... Layout Analysis, **Machine** Learning, Decision tree, Support **Vector Machine**, **...**deeper **language** analysis for both table **detection** and interpretation. **...**

The example in (21) presents more articulate asymmetric relations, as well as it illustrates a situation similar to the one related to the example (16). The example in (22) includes a *by*-phrase, which is an adjunct related to the external argument of the predicate *detection*. The natural language operator BY introduces this adjunct. This operator, like the OF operator, does not have the symmetric properties of the Boolean operators AND and OR. BY introduces an asymmetric relation between a predicate and an adjunct. This is represented by the structure in (23), where the by-phrase is linked to the internal argument of the nominal predicate.

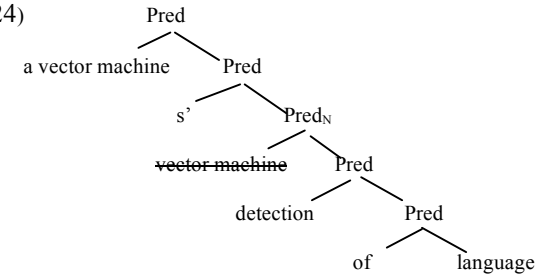
(23)



As it is the case for the example in (16), a search engine oriented by the recovery of asymmetric relations will correctly identify the external argument of nominal expressions with different overt distribution of arguments. In the case at hand, and given the properties of English, the structure in (23) and the structure in (24), with the possessive operator S', will be analyzed as equivalents with respect to the identification of the

external argument. In both cases the natural language operator projection is linked to the external argument.

(24)



Precision and recall benefit from an asymmetry-oriented search engine based on the recovery of fine-grained linguistic knowledge, including argument structure asymmetries.

IV. QUESTION ANSWERING SYSTEMS

A question answering system takes requests for information on the part of users 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 [30] for a discussion of the characteristics of each type of systems).

These two types of QASs can be distinguished in terms of their knowledge bases. The knowledge base 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 a new type of question answering applications after the advent of the World Wide Web in the early 90s, 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 [31]) uses the web only to extract information which they then store 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 Di Sciullo and Aguero [30] for discussion of web-based question answering systems.

Question answering systems are also based on poor linguistic knowledge. Even the MIT START system (Katz [31]), which uses inverse transformations in the processing of questions, still failed to provide answers to simple forms of questions. While a correct answer to questions in (25) is provided by START, no answer is available for the questions in (26) and (27).

(25) a. Q: What did Wittgenstein wrote?
b. Q: What did Wittgenstein do?

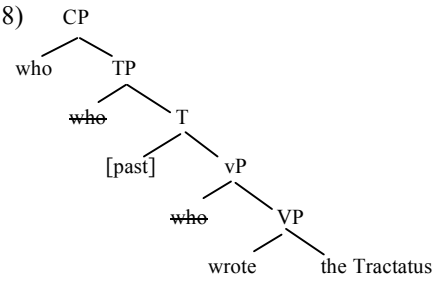
(26) a. Q: Who wrote the Tractatus?
A: Sorry, I don't know the answer.
b. Q: Who is the author of the Tractatus?
A: I don't known.

(27) a. Q: Where was the Tractatus written?
b. Q: When was the Tractatus written?

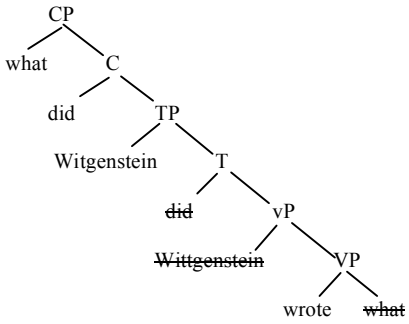
The form of the questions in (25) differs from the ones in (26) and (27). It is the internal argument that is questioned in (25), whereas it is the external argument in (26) and an adjunct in (27). These results clearly show that the asymmetry between the internal and the external arguments, as well as the asymmetry between arguments and the adjuncts are not taken into consideration by the question answering system.

In languages such as English, as well as in most languages of the world, see Di Sciullo [8], the morphological form of the interrogative pronoun indicates the role of the questioned argument or the adjunct with respect to the predicate. The interrogative pronouns *who* and *what* are limited to questioning the arguments of a predicate, whereas the pronouns *where* and *when* are limited to questioning spatial and temporal modification of the event denoted by a predicate. The natural language identification of the properties of the referent targeted by a question is obviously not taken into consideration by the question answering system. If it were, the system would offer an answer to the question in (26a), since it did answer to the question in (25a). The success obtained with respect to the questions in (25) is not due to the actual understanding of questions, but rather to the use of proximity calculi and pattern matching methods. Thus, in order to understand the question in (25a), the system should be able to deduce that *Wittgenstein* is the external argument (agent of the event denoted by the predicate) and that the internal argument of this event is the Tractatus. In order to understand the question in (26), the system should be able to identify the external argument (agent of the event denoted by the predicate *wrote the Tractatus*). The configurational asymmetry between questioning the internal argument and questioning the external argument is represented in (28) and (29). In (28), it is the external argument of the predicate that is questioned, as can be seen by the fact that the interrogative pronoun origins in the specifier of the verbal predicate projection. In (29), it is the internal argument of the predicate that is questioned, as can be seen by the fact that the interrogative pronoun origins in the complement position of the verbal predicate.

(28)



(29)



The identification of the asymmetric properties of the arguments with respect to a predicate is an important aspect of a questioning system oriented by the recovery of asymmetric relations. Because such system is based on the processing of natural language asymmetries, the processing of questions is just a sub-case of the more general approach in information processing.

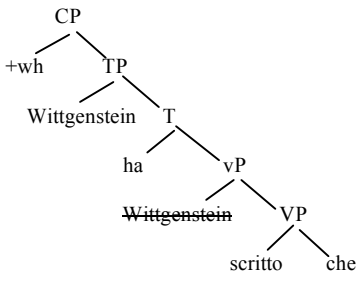
Including rich linguistic knowledge in natural language processing allows upgrading the performance of question answering systems and contributes to natural language understanding. The structural asymmetry of the arguments of a predicate are constant across languages and make the interpretation of questions possible notwithstanding the overt occurrences of the interrogative pronoun or the overt form of the question.

Thus, in languages such as Italian, an interrogative pronoun may occupy the canonical internal argument position, as the examples in (30) illustrate. This is not the case in English, where an interrogative pronoun must be in the left periphery of a question.

(30) Wittgenstein ha scritto che? (Italian)
(Wittgenstein has written what)
What did Wittgenstein write?'

This property of languages such as Italian is directly processed by a question answering system based on the recovery of asymmetric relations. The Italian interrogative pronoun *che* 'what' occupies the canonical argument position overtly and the +wh Operator feature is in C.

(31)



Languages such as English also differ from other languages, including Italian, with respect to the possibility of stranding the preposition in a question including a prepositional constituent, as in the example in (32). START provides no answer to this question.

(32) What is the Tractatus about?

The recovery of the asymmetric relations is necessary for their understanding and thus for their semantic interpretation. The referent of a question is a set of individuals; the properties of natural languages restrict the sets of individuals, which may qualify as possible answers to the question. The internal /external argument asymmetry brought about by the morphological form of the interrogative pronoun and the canonical argument position occupied by the pronoun restricts the choice of the set of individuals. The inclusion of fine-grained linguistic knowledge in the processing of the syntactic form of questions contributes to the efficiency of question answering systems. In a system oriented by the recovery of asymmetric relations the variation in the form of questions is however parasitic on the canonical position of the arguments, thus ensuring the preservation of the meaning of questions whatever language is used.

V. CONSEQUENCE FOR INFORMATION PROCESSNG

NLP cannot be reduced to the tagging of the words in documents. The asymmetric relations, couched in terms of structural relations, must be recovered in order to identify the semantics of linguistic expressions, and thus the referent of linguistic expressions.

For the viewpoint of asymmetry based NLP, it is possible to consider the development of a "natural" semantic web based on natural language syntax-semantics, which is a prerequisite to any semantic web designed to process large scale real-world semantic knowledge. (Berners-Lee, Hendler, and Lassila [32]).

NLP based on the recovery of asymmetric relations may lead to the development of an information-processing model integrating the form and the semantic interpretation of linguistic expressions, notwithstanding language diversity. Information processing systems oriented by the recovery of asymmetric relations will contributes to the development of efficient software that will process information supported by language efficiently, as they will be based on the modelling of a core property of natural language.

ACKNOWLEDGMENT

This work is supported in part by funding from the Social Sciences and Humanities Research Council of Canada to the Major Collaborative Research on Interface Asymmetries, grant number 214-2003-1003, awarded to professor Anna Maria Di Sciullo, of the Département de Linguistique at Université du Québec à Montréal. www.interfaceasymmetry.uqam.ca

REFERENCES

[1] N. Chomsky, "Minimalist inquiries: the framework". In *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, eds. Roger

Martin, David Michaels and Juan Uriagereka, 89-155. Cambridge, Mass: MIT Press, 2000.

[2] N. Chomsky, "On Phases". In *Foundational Issues in Linguistic Theory. Essays in Honor of Jean-Roger Vergnaud*, eds. Robert Freidin, Carlos Peregrín Otero and Maria Luisa Zubizarreta, 133–166. Cambridge, MA: MIT Press, 2008.

[3] A. M. Di Sciullo and E. Williams, *On the Definition of Word*, Cambridge, Mass.: The MIT Press, 1987.

[4] R. Kayne, *The Antisymmetry of Syntax*, Cambridge, Mass.: The MIT Press, 1994.

[5] R. Kayne, Why are there no head-directionality parameters, in M. Byram Washburn, K. McKinney- Bock, E. Varis, A. Sawyer and B. Tomaszewicz (eds.) *Proceedings of the 28th West Coast Conference on Formal Linguistics*, Cascadilla Proceedings Project, Somerville, MA, 1-23, 2011.

[6] A. Moro, *Dynamic Antisymmetry*, Cambridge, Mass.: The MIT Press, 2000.

[7] A. Moro, The Boundaries of Babel, *The Brain and the Enigma of Impossible Languages*, Cambridge, Mass: The MIT Press, 2008.

[8] K. Hale and S. J. Keyser, *Prolegomenon to a Theory of Argument Structure*, Cambridge, Mass.: The MIT Press, 2002.

[9] H. van der Hulst and N. Ritter, "Head-driven Phonology", in H.G. van der Hulst and N. Ritter (eds.) *The Syllable: Views and Facts*, Berlin: Mouton de Gruyter, 1999, pp. 113-169.

[10] H. van der Hulst and N. Ritter, "Levels, Constraints and Heads", in A.M. Di Sciullo (ed.) *Asymmetry in Grammar, volume 2: Morphology, Phonology and Acquisition*, Amsterdam: John Benjamins, 2003, pp. 151-193.

[11] E. Raimy, "Remarks on Backcopying", Linguistic Inquiry, Vol. 31, 2000, No 3, pp. 541-552.

[12] E. Raimy, "Asymmetry and Linearization in Phonology", in A. M. Di Sciullo (ed.) *Asymmetry in Grammar, volume 2: Morphology, Phonology and Acquisition*, Amsterdam: John Benjamins, 2003, pp. 133-150.

[13] A. M. Di Sciullo, *Asymmetry in Morphology*, Cambridge, Mass.: The MIT Press, 2005.

[14] A. M. Di Sciullo, "Perspectives on Morphological Complexity", in *Morphology. (Ir)regularity, Frequency, Typology*, F. Kiefer, M. Ladanyi and P. Siptar (dir.) .Amsterdam : John Benjamins, 2012, pp.105-135.

[15] A. M. Di Sciullo, "I-Morphology, Operations, Interfaces and Complexity", in P. Kosta, S. Franks and L Schürcks (ed,) *Minimalism and Beyond: Radicalizing the interfaces,* Language Faculty and Beyond, Amsterdam: John Benjamins. 2013.

[16] R. Wall, *Introduction to Mathematical Linguistics*, New Jersey: Prentice Hall, 1972.

[17] A.M. Di Sciullo, I-Morphology, Operations, Interfaces and Complexity. In P. Kosta, S. Franks and L Schürcks (ed,) *Minimalism and Beyond: Radicalizing the interfaces.* Language Faculty and Beyond. Amsterdam: John Benjamins, 2013.

[18] M. Marcus, A Theory of Syntactic Recognition for Natural Language. M. Marcus, *A Theory of Syntactic Recognition for Natural Language*, Cambridge, Mass.: The MIT Press, 1980.

[19] R. Berwick and A. Weinberg, *The Grammatical Basis of Linguistic Performance*, Cambridge, Mass.: The MIT Press, 1984.

[20] R. Berwick, *The Acquisition of Syntactic Knowledge,* The MIT Press, 1985.

[21] R. Berwick, S. Abney, and C. Tenny (eds.), *Principle-Based Parsing: Computation and Psycholinguistics*, Studies in Linguistics and Philosophy, Dordrecht: Kluwer, 1991.

[22] S. Fong, Computational Properties of Principle-Based Grammatical Theories, Doctoral Dissertation, Artificial Intelligence Laboratory, MIT, 1991.

[23] A. M. Di Sciullo, "Parsing Asymmetries", *Natural Language Processing,* Springler Computer Science Press, 2000, pp. 24-39.

[24] A. M. Di Sciullo and S. Fong, "Morpho-Syntax Parsing", in A. M. Di Sciullo (ed.) *UG and External Systems. Language, Brain and Computation,* Amsterdam: John Benjamins, 2005, pp. 247-268.

[25] S. Fong, "Computation with Probes and Goals", in A.M. Di Sciullo (ed.) *UG and External Systems. Language, Brain and Computation,* Amsterdam: John Benjamins, 2005, pp. 311-334.

[26] H. Harkema, "Minimalist Languages and the Correct Prefix Property", in A. M. Di Sciullo (ed.) *UG and External Systems. Language, Brain and Computation,* Amsterdam: John Benjamins, 2005, pp. 289-310.

[27] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[28] T. Strzalkowski (ed.) *Natural Language Information Retrieval*, Dordrecht: Kluwer, 1999.

[29] W. B. Frakes and R. Baeza-Yates, *Information Retrieval*, Prentice Hall, 1992.

[30] A. M. Di Sciullo and C. Aguero, "Natural Language Asymmetries and the Construction of Question Answering Systems", *Proceedings of the 7th World Multiconference on Systemics, Cybernetics, and Informatics*, Vol. 1, 2003, pp. 13-17.

[31] B. Katz, "From Language Processing to Information Access on the World Wide Web", *Papers from the 1997 AAAI Symposium*, 1997, pp 77-86.

[32] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", *Scientific American*, Vol 284, no. 5, 2001, pp 34-43.