Biolinguistics, Natural Language Processing, and Digital Libraries

Anna Maria Di Sciullo
Professor, Département de Linguistique
Université du Québec à Montréal
405 Rue Sainte-Catherine Est, Montréal, QC H2L 2C4
(E): di_sciullo.anne-marie@uqam.ca

DOI: 10.18329/09757597/2015/8101

World Digital Libraries 8(1): 1–18 (2015)

Abstract

Notwithstanding the progress achieved in the conception and the implementation of Digital Libraries, further research is needed to improve their efficiency. In this article, the author identifies shortcomings of current Digital Libraries and present the basic elements of a biolinguistic approach to natural language processing with consequences for development of more efficient information systems in the future.

Keywords: Biolinguistic approach, Linguistic expressions, Lexical projections, Conceptual-intentional systems, Dspace, Asymmetric relationships

1. A Biolinguistic Approach to Digital Libraries

Digital Libraries are network information systems supporting search and retrieval of items from structured collections. Their purpose is to enable users to interact effectively with information distributed across a network. Schatz (1997) outlined their historical evolution starting from the 1960s with the retrieval of scientific literature from bibliographic databases, evolving into full-text retrieval and finally into a document search on the Internet (Figure 1).

'In the historical evolution of digital libraries, the mechanisms for retrieval of scientific literature have been particularly important. Grand visions in 1960 led first to the development of text search, from bibliographic databases to full-text retrieval. Next, research prototypes catalysed the rise of document search, from multimedia browsing across local-area networks to distributed search on the Internet. By 2010, the visions will be realized, with concept search enabling semantic retrieval across large collections' (Schatz 1997: 327).

According to Schatz (1997), concept search, enabling semantic retrieval across large collections, were to be achieved by 2010. However, while progress has been achieved with respect to the modelling and the implementation of Digital Libraries, for example, the DELOS Reference Model (Candela *et al.* 2008, 2011), semantic retrieval across large collections is yet to be completed. The efficiency of Digital Libraries is not optimal and further work is needed to improve the search and retrieval of scientific literature.

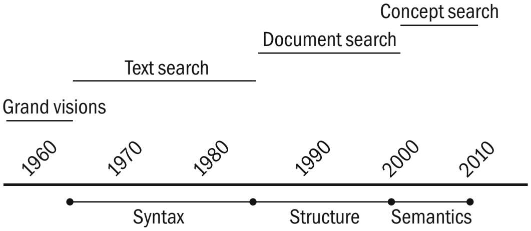


Figure 1: Historical evolution starting from the 1960s with the retrieval of scientific literature from bibliographic databases
Source: Schatz (1997).

The first problem I address, through this article, is—What are the principles of efficient computation that can be imposed on information systems such as Digital Libraries, in order to narrow down their search space and retrieve relevant items? I argue in favour of a model of Digital Libraries, based on the properties of the Language Faculty and principles of efficient computation.

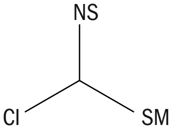
The investigation of the biological properties of the Language Faculty is the object of inquiry of the interdisciplinary field of Biolinguistics, which brings together biology, linguistics, psychology, neurosciences, and computer sciences (Lenneberg 1967; Jenkins 2000, 2004; Chomsky 2005, 2008, 2013; Di Sciullo 2011, Di Sciullo et al. 2010b; Di Sciullo and Boeckx 2011, a.o.). I thus formulate my proposal as follows:

(1) *The Biolinguistic Approach to Digital Libraries:* Efficient Digital Libraries incorporate Natural Language Processing systems based on the properties of the Language Faculty.

Biolinguistics aims to further understand the biological basis of language, that is, the properties of the Language Faculty. The Biolinguistic Approach to Digital Libraries addresses the problem of the search and retrieval of documents from a repository on the basis of recent advances in the properties of the computational procedure of the Language Faculty and the principles of efficient computation. This computational procedure is efficient as it makes humans capable of developing a grammar on the basis of impoverished experience and use language creatively to express simple and complex thoughts. The incorporation of the properties of the Language Faculty and the principles of efficient computation in Natural Language Processing systems may lead to more efficient information systems, including Digital Libraries.

Current information retrieval systems analyse linguistic expressions in terms of flat structure and statistical calculi.[1] Current linguistic theory, however analyses linguistic expressions in terms of hierarchical structures generated by a small set of operations and economy principles. Recent works in linguistic theory (Chomsky 1995, 2001, 2005, 2013, 2014; Epstein and Seely 2002; Hauser, Chomsky and Fitch 2002; Kayne 1994, 2010; Moro 2000, 2010; Di Sciullo 2005a, 2011, 2014, a.o.) develop the hypothesis that the narrow syntactic component of the Language Faculty satisfies conditions of highly efficient computation. Thus, the language faculty could be close to an optimal solution to the problem of linking forms perceived by the sensorimotor system (SM) and meanings computed by the conceptual-intentional (CI) systems via the computations of the narrow syntax component (NS). In other words, the language system may provide a near optimal solution that satisfies the interface conditions of interpretability.

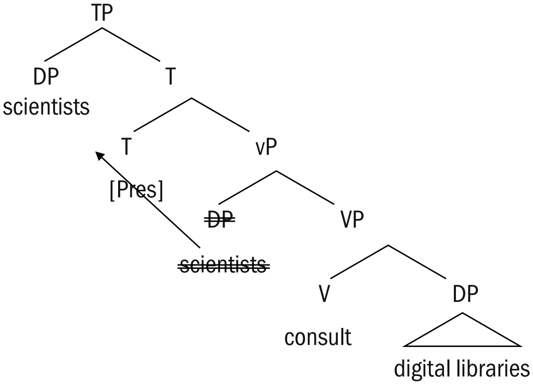
(2)



In the Minimalist Program (Chomsky 1995–2014) the core generative procedure of the Language property is reduced to the dyadic and recursive operation Merge. This operation applies to two syntactic objects and derives a new syntactic object.[2]

The result of the recursive application of this operation may be represented by hierarchical structures, as in (3a) or by a parenthetical expression, as in (3b) for the expression, *scientists consult digital libraries.*

(3)



$[_{TP}$ DP T $[_{vP}$ ~~DP~~ v $[_{VP}$ V DP]]]

The representations in (3) include functional categories such as Tense (T), Determiners (D), and their phrasal projections, respectively, TP and DP, in addition to lexical categories such as Verb (V) and Nouns (N) and their projections VP and NP. Thus, in addition to lexical projections, linguistic expressions include functional projections, which also spell out the semantic relations between linguistic constituents. These projections are crucial for the CI interpretation. For example, the structures in (3) include a TP, which is part of the structure of all sentences and provides the temporal interpretation of linguistic expressions. However, the T head may be spelt out by morphological material in certain languages, as is the case in the Romance languages, but not in others, as it is generally the case in Chinese for instance. The DP subject generated in the VP is displaced in a higher position than T. Displacement or Remerge is the application of Merge to a constituent that has already been merged in a previous step of the derivation. This operation leaves a copy in the position from which the movement takes place (represented by strikethrough in [3]). This simple generative procedure enables the processing of complex thoughts expressed by language. The hierarchical structure generated by grammar is interpreted by the semantic rule of *Functional Application,* (4), which applies at each step of the derivation and ensures efficient semantic interpretation of the linguistic expressions.

(4) Functional Application: If α is a branching node and $\{\beta,\gamma\}$ the set of its daughters, then, for any assignment g, $||\alpha||^g = ||\beta||^g(||\gamma||^g)$. (Heim and Kratzer 1998)

The computational efficiency should be part of information systems (Karamanis *et al.* 2007; Ahler *et al.* 2007; Neveol *et al.* 2007; Yu and Kaufman 2007; Di Sciullo 2014, a.o.). Most information retrieval systems however do not incorporate a grammar that is capable of processing natural language efficiently (Buettcher *et al.* 2010, a.o.). Finite State Grammars (FSG) are used in Information Retrieval systems for spell-checkers, morphological stemmers, and partial parsing. However, FSG derive flat structures not hierarchical structures and their generative capacity cannot fully describe the properties of linguistic expressions.[3]

The capacity to process hierarchical structures is specific to the human species. Comparative evolutionary studies (Fitch and Hauser 2004; Murphy *et al.* 2008) indicate that birds and non-human primates can compute a first-degree FSG, where elements in a string of sounds have specific orders, each predicted by simple statistical association, but not abstract hierarchical structures and complex dependencies observed in linguistic expressions. The results of Fitch and Hauser (2004) experiments show that with cotton-top tamarin monkeys have the capacity to learn artificial languages derived by a FSG, sequences isomorphic to *ababab,* but not artificial grammar derived by phrase structure grammar (CFG), structures isomorphic to *aaabbb* (Figure 2). Thus, contrary to humans, non-human primates cannot process abstract projections and nested dependencies.

Building on Fitch and Hauser's (2004) results, Friederici's (2009) fMRI localized specific areas in the human brain BA_{44}, $BA_{45}B$ for processing hierarchical structure. Interestingly, even though these areas are also observed in the macaque brain, their size and granularity are reduced (Figure 3)

> '[. . .] the human ability to process hierarchical structures may depend on the brain region which is not fully developed in monkeys but is fully developed in humans, and that this phylogenetically younger piece of cortex may be fundamentally relevant for the learning of the PSG' (Friederici 2009: 185).

Friederici's findings lend biological (neuro-anatomical) support to the hierarchical structure derived by the operations of the Language Faculty, as opposed to the kind of representations derived by animal computations. Summarizing so far, in the Biolinguistic Approach to Digital Libraries, FSG cannot process efficiently the documents populating

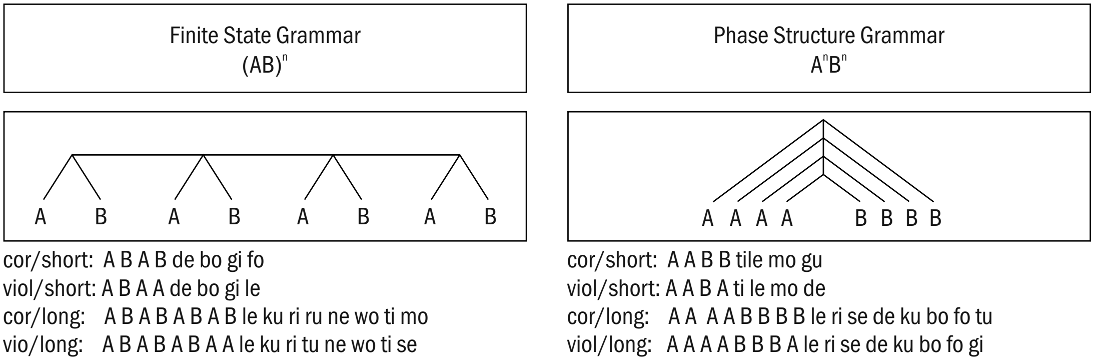
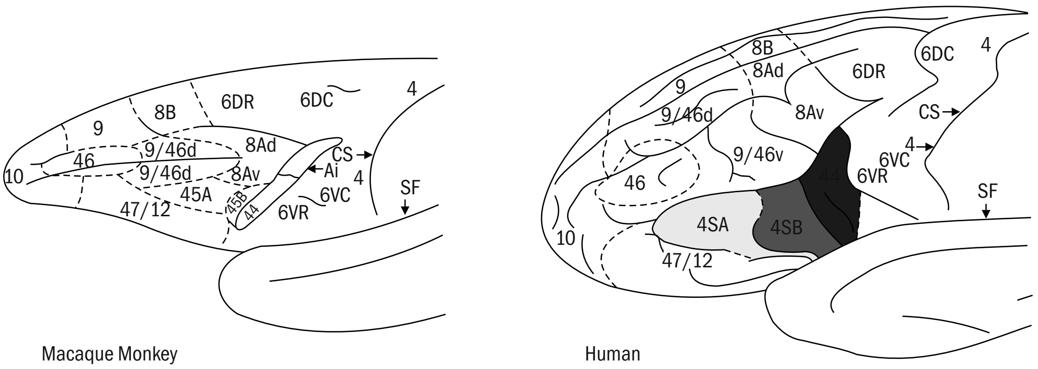

Finite State Grammar
(AB)ⁿ
A B A B A B A B
cor/short: A B A B de bo gi fo
viol/short: A B A A de bo gi le
cor/long: A B A B A B A B le ku ri ru ne wo ti mo
vio/long: A B A B A B A A le ku ri tu ne wo ti se
Phase Structure Grammar
AⁿBⁿ
A A A A B B B B
cor/short: A A B B tile mo gu
viol/short: A A B A ti le mo de
cor/long: A A A A B B B B le ri se de ku bo fo tu
viol/long: A A A A B B B A le ri se de ku bo fo gi

Figure 2: *(Please refer to the text above)*
Source: Adapted from Friederici *et al.* (2006)



Macaque Monkey

Human

Figure 3: *(Please refer to the text)*
Source: Adapted from Petrides and Pandya (1994)

the collections. Instead, the query and the documents of the collections can be analysed in terms of grammar with a higher generative capacity, capable of deriving binary branching hierarchical structures including, functional categories. These hierarchical structures articulate the syntax-semantic relations between the constituents of linguistic expressions.

2. Operating Digital Libraries, Two Examples

Digital Libraries are searchable repositories of scientific material and their efficiency depends on the progress achieved in information systems. Current digital research libraries are operating systems extracting information on the basis of keyword search and statistical methods, such as the so-called ‘bag of words’ method, as well as other similarity metrics for character and string matching. The Biolinguistic Approach to Digital Libraries predicts that information systems that do not rely on the core properties of the Language Faculty will not be optimal. This prediction is borne out as I illustrate by testing the performance of two online digital libraries—DSpace@MIT and PubMed.

2.1 DSpace@MIT

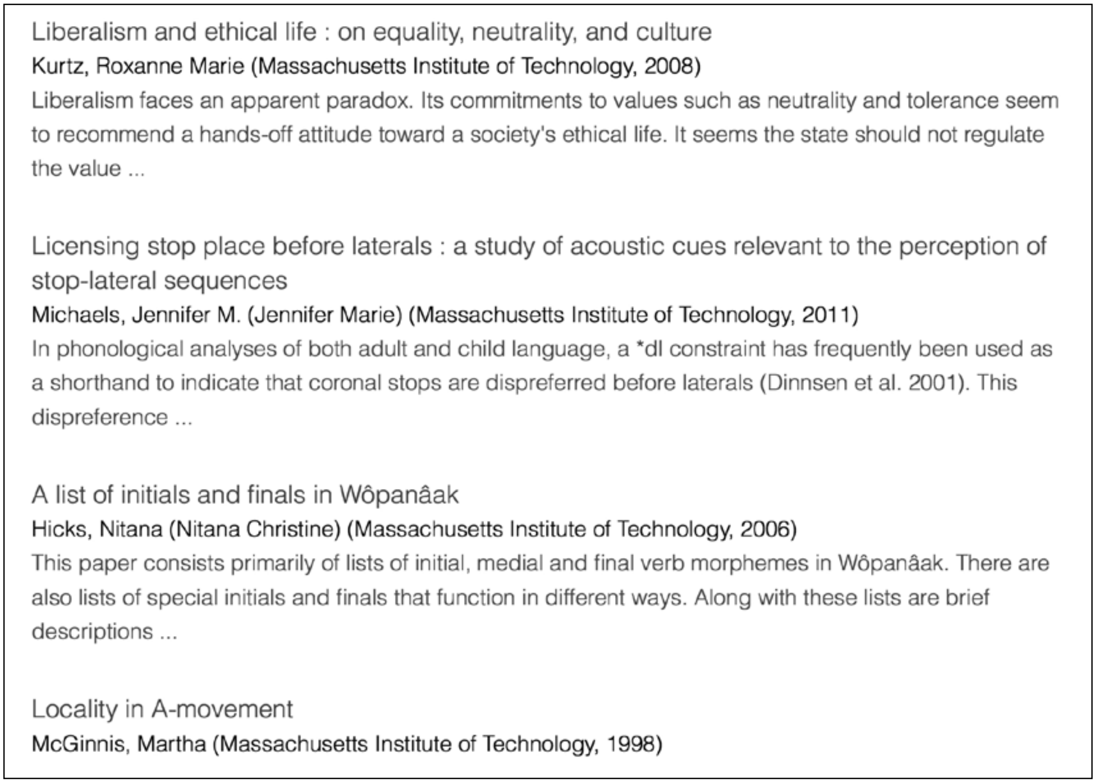
DSpace@MIT is a searchable repertoire of Massachusetts Institute of Technology (MIT) dissertations. DSpace was jointly developed by Hewlett Packard Labs and MIT in order to create an open source software solution for archiving digital content. The first release of the software goes back to 2002. It is an open source technology for global communities who manage, preserve, and provide access to digital content. The metadata, including access and configuration information is stored in a relational database and supports the use of PostgreSQL and Oracle database. DSpace uses standard Dublin core descriptive metadata (keywords, descriptions) to aid search and retrieval. All metadata and text is indexed and fully searchable. The system can customize specific fields to enable browsing. It can also choose what fields and text to be indexed for search. See www.DSpace.org for further specifications. DSpace@MIT browses MIT dissertations by issue date, authors, titles, subjects, series, and the ISSN/ISBN numbers. The types of search are divided into basic and advanced. The basic search is typically based on keywords and the search types include the following:

(5) a. Keyword
b. Title begins with . . .
c. Title keyword
d. Author (last name first)
e. Author keyword

f. Call number begins with . . .
g. Subject begins with . . .
h. Subject keyword
i. Series title begins with . . .
j. Series title keyword
k. ISSN/ISBN

The search interface specifies the restrictions imposed on the syntax of the query. As it is usual in the keyword-based approach to information retrieval, stop words (functional categories such as articles and prepositions) need to be omitted from the query. In effect if (5b) is chosen, the system requires to omit initial articles such as 'a', 'an', 'the'. These elements are however an essential part of the syntax-semantic relation and affect the interpretation of linguistic expressions. The system will not search on a stop word in a keyword search unless it is entered in quotation marks. For example, to carry out a title keyword search using the terms *set theory*, enter your search as 'set' 'theory'. This restriction on the syntax of the query is designed to comply with the properties of the information retrieval system, which is based on keywords search, that is, an automated search based on the matching of strings of characters and designated sequences of words. These restrictions, while they comply with the query of the language, do not lead to efficient retrieval, as can be seen by performing the following simple tests. Let us browse through the Department of Linguistics and Philosophy dissertations by title, submit queries to the system, and evaluate the results. A one word query (6), and two multiword queries (7) and (8) are used for this simple test. The results are not optimal, as can be seen by the first four dissertation titles and short abstracts retrieved by the system.

(6) DSpace@MIT results for the query: lexicon



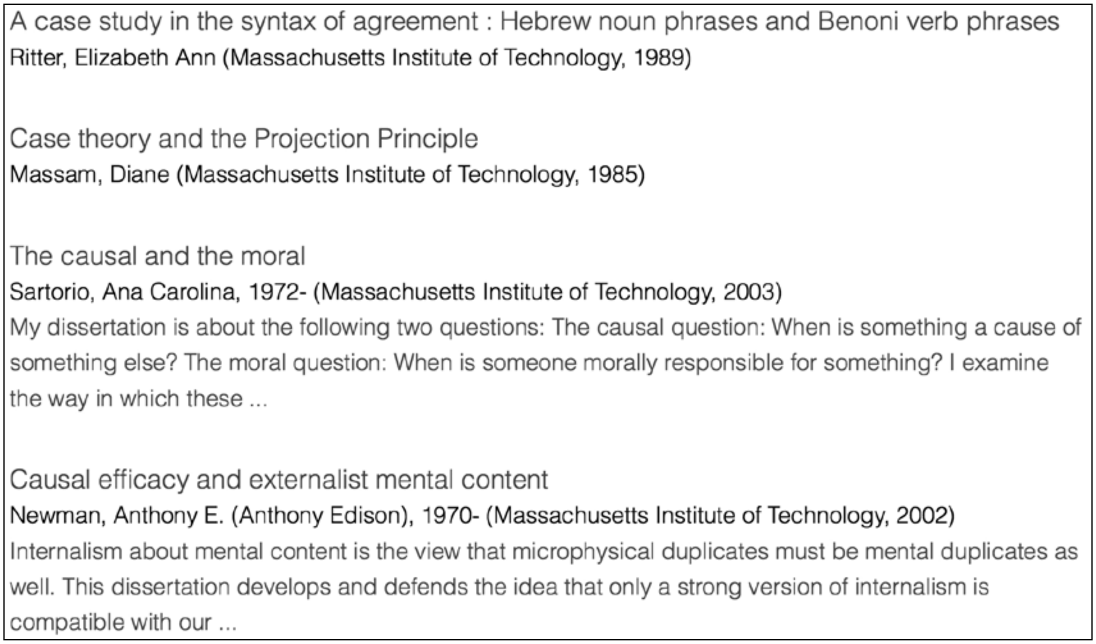
Liberalism and ethical life : on equality, neutrality, and culture
Kurtz, Roxanne Marie (Massachusetts Institute of Technology, 2008)
Liberalism faces an apparent paradox. Its commitments to values such as neutrality and tolerance seem to recommend a hands-off attitude toward a society's ethical life. It seems the state should not regulate the value ...

Licensing stop place before laterals : a study of acoustic cues relevant to the perception of stop-lateral sequences
Michaels, Jennifer M. (Jennifer Marie) (Massachusetts Institute of Technology, 2011)
In phonological analyses of both adult and child language, a *dl constraint has frequently been used as a shorthand to indicate that coronal stops are dispreferred before laterals (Dinnsen et al. 2001). This dispreference ...

A list of initials and finals in Wôpanâak
Hicks, Nitana (Nitana Christine) (Massachusetts Institute of Technology, 2006)
This paper consists primarily of lists of initial, medial and final verb morphemes in Wôpanâak. There are also lists of special initials and finals that function in different ways. Along with these lists are brief descriptions ...

Locality in A-movement
McGinnis, Martha (Massachusetts Institute of Technology, 1998)

(7) DSpace@MIT results for the query: Case projection principle



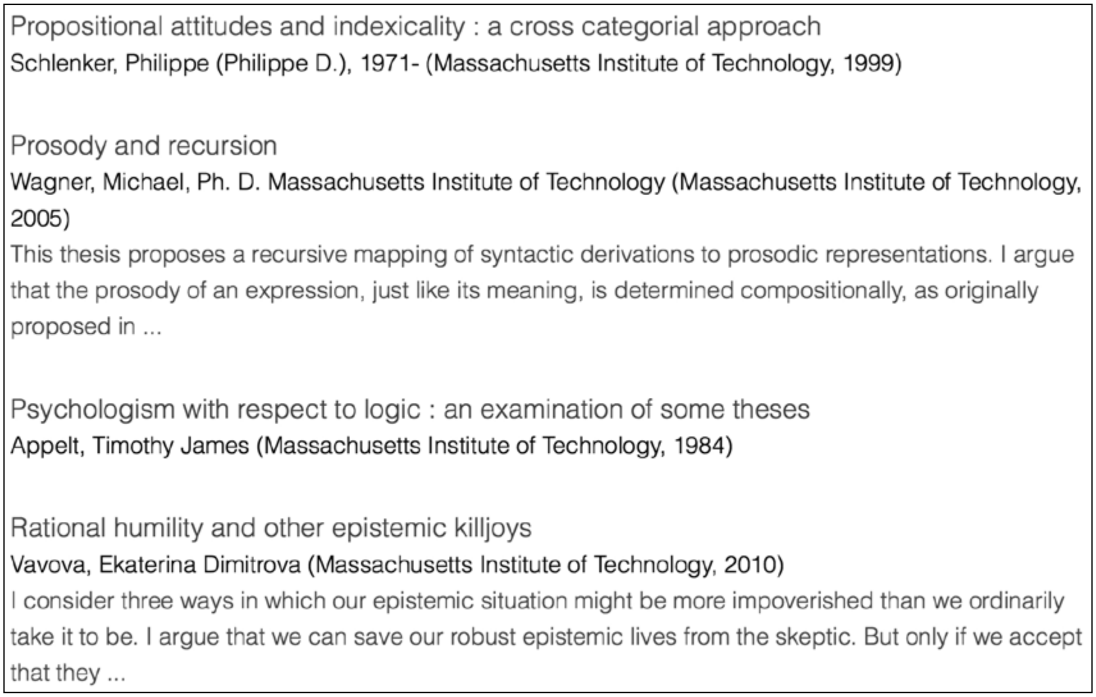
A case study in the syntax of agreement : Hebrew noun phrases and Benoni verb phrases
Ritter, Elizabeth Ann (Massachusetts Institute of Technology, 1989)

Case theory and the Projection Principle
Massam, Diane (Massachusetts Institute of Technology, 1985)

The causal and the moral
Sartorio, Ana Carolina, 1972- (Massachusetts Institute of Technology, 2003)
My dissertation is about the following two questions: The causal question: When is something a cause of something else? The moral question: When is someone morally responsible for something? I examine the way in which these ...

Causal efficacy and externalist mental content
Newman, Anthony E. (Anthony Edison), 1970- (Massachusetts Institute of Technology, 2002)
Internalism about mental content is the view that microphysical duplicates must be mental duplicates as well. This dissertation develops and defends the idea that only a strong version of internalism is compatible with our ...

(8) DSpace@MIT results for the query: Prolegomena to a theory of word formation



Propositional attitudes and indexicality : a cross categorial approach
Schlenker, Philippe (Philippe D.), 1971- (Massachusetts Institute of Technology, 1999)

Prosody and recursion
Wagner, Michael, Ph. D. Massachusetts Institute of Technology (Massachusetts Institute of Technology, 2005)
This thesis proposes a recursive mapping of syntactic derivations to prosodic representations. I argue that the prosody of an expression, just like its meaning, is determined compositionally, as originally proposed in ...

Psychologism with respect to logic : an examination of some theses
Appelt, Timothy James (Massachusetts Institute of Technology, 1984)

Rational humility and other epistemic killjoys
Vavova, Ekaterina Dimitrova (Massachusetts Institute of Technology, 2010)
I consider three ways in which our epistemic situation might be more impoverished than we ordinarily take it to be. I argue that we can save our robust epistemic lives from the skeptic. But only if we accept that they ...

DSpace@MIT brings back no relevant results for the first search, which restricts the search space to the MIT Department of Linguistic and Philosophy community and targets the title of a dissertation including the word, lexicon. For the second, more complex search targeting dissertations on case projection principle, a relevant result is retrieved by the system, although it is ranked in second position while the first document retrieved is not a dissertation on the topic. In the last case, several titles were retrieved by the search, however, none of them matched the terms of the query. The system always returns documents even when there are no relevant documents to be retrieved, as it is the case for this last query, which is the title of an article published in *Linguistic Inquiry*, an MIT Press Journal, and not a title of an MIT dissertation. In sum, DSpace@MIT retrieves dissertations from the MIT repository; however, all the documents retrieved are not necessarily relevant. The lack of precision of DSpace@MIT resides mainly in the fact that the Information Retrieval system used does not rely on the syntax-semantic analysis of the query and of the documents in the targeted collection.

2.2 PubMed

PubMed is an open access database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. PubMed comprises more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. The strategy to search PubMed for simple subjects is the following:

(9) Simple subject search strategy

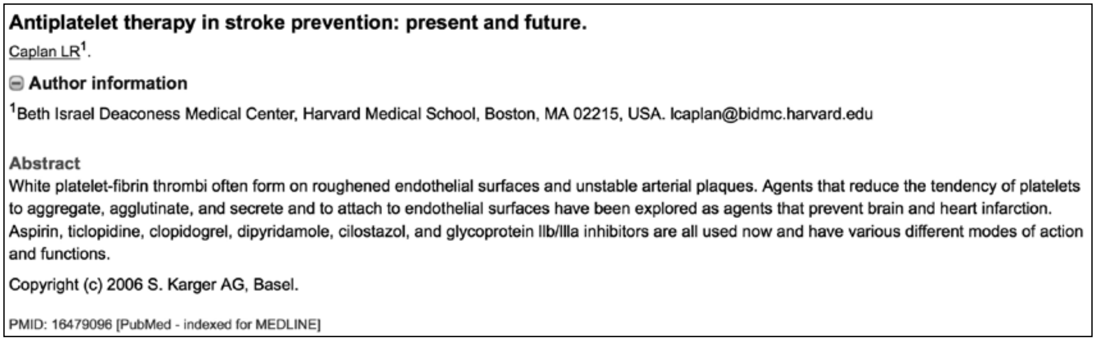
1. Identify the key concepts for your search.
2. Enter the terms (or key concepts) in the search box.
3. Suggestions will display as you type your search terms.
4. Click Search.

Here again, the search engine is based on keywords. Thus, no punctuation, tags or operators should be used in the query. A PubMed search for articles on the use of aspirin for heart attack prevention brings back documents where the substantive terms of the query are distributed in different sentences, as (10) reveals.

(10) PubMed result for the query on the use of aspirin for heart attack prevention

The strategy to use PubMed for complex subjects is mediated by the use of Boolean operators AND, OR and NOT, as the following queries illustrate:

(11) a. (use aspirin) AND (heart attack prevention)



Antiplatelet therapy in stroke prevention: present and future.

Caplan LR[1].

⊖ **Author information**

[1]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02215, USA. lcaplan@bidmc.harvard.edu

Abstract

White platelet-fibrin thrombi often form on roughened endothelial surfaces and unstable arterial plaques. Agents that reduce the tendency of platelets to aggregate, agglutinate, and secrete and to attach to endothelial surfaces have been explored as agents that prevent brain and heart infarction. Aspirin, ticlopidine, clopidogrel, dipyridamole, cilostazol, and glycoprotein IIb/IIIa inhibitors are all used now and have various different modes of action and functions.

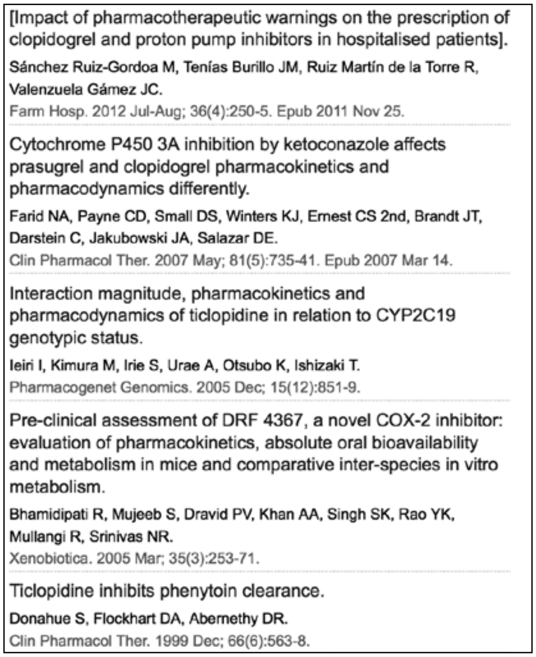
Copyright (c) 2006 S. Karger AG, Basel.

PMID: 16479096 [PubMed - indexed for MEDLINE]

b. (sickle-cell anemia) AND ((Genetic Counselling OR Inheritance pattern AND genetics))
c. (inhalation therapy pneumonia) AND systematic[sb]

However, even by using Boolean operators, the results are not optimal. The results of PubMed Clinical queries do not systematically satisfy the request for information formulated by the query. For the query whether Ticlopidine is a potent inhibitor for CYP2C19, the following five documents are retrieved. Only two out of the five documents retrieved are relevant. Moreover, the system does not rank the documents retrieved in an accurate order of relevance.

(12) PubMed Clinical Queries results for the query whether Ticlopidine is a potent inhibitor for CYP2C19



[Impact of pharmacotherapeutic warnings on the prescription of clopidogrel and proton pump inhibitors in hospitalised patients].
Sánchez Ruiz-Gordoa M, Tenías Burillo JM, Ruiz Martín de la Torre R, Valenzuela Gámez JC.
Farm Hosp. 2012 Jul-Aug; 36(4):250-5. Epub 2011 Nov 25.

Cytochrome P450 3A inhibition by ketoconazole affects prasugrel and clopidogrel pharmacokinetics and pharmacodynamics differently.
Farid NA, Payne CD, Small DS, Winters KJ, Ernest CS 2nd, Brandt JT, Darstein C, Jakubowski JA, Salazar DE.
Clin Pharmacol Ther. 2007 May; 81(5):735-41. Epub 2007 Mar 14.

Interaction magnitude, pharmacokinetics and pharmacodynamics of ticlopidine in relation to CYP2C19 genotypic status.
Ieiri I, Kimura M, Irie S, Urae A, Otsubo K, Ishizaki T.
Pharmacogenet Genomics. 2005 Dec; 15(12):851-9.

Pre-clinical assessment of DRF 4367, a novel COX-2 inhibitor: evaluation of pharmacokinetics, absolute oral bioavailability and metabolism in mice and comparative inter-species in vitro metabolism.
Bhamidipati R, Mujeeb S, Dravid PV, Khan AA, Singh SK, Rao YK, Mullangi R, Srinivas NR.
Xenobiotica. 2005 Mar; 35(3):253-71.

Ticlopidine inhibits phenytoin clearance.
Donahue S, Flockhart DA, Abernethy DR.
Clin Pharmacol Ther. 1999 Dec; 66(6):563-8.

These results illustrate the fact that PubMed does not rely on an optimal information retrieval system. Like DSpace@MIT, the information retrieval system is based on keyword search and thus it requires the use of a special syntax for the query, including the use of Boolean operators, and excluding the use of functional words in the query. Instead of imposing syntactic restrictions on the formulation of a query by the user, an information retrieval system should instead be able to process queries formulated in the natural language. The information systems based on the syntax-semantic analysis of the query and documents that populate the collections should achieve higher levels of performance than systems based on keyword search.

Summarizing, both DSpace@MIT and PubMed are Digital Libraries enabling search and retrieval of material from scientific collections. Both information systems operate on keyword search and statistical methods are used to identify and rank the documents satisfying the request of information formulated by the query. Simple tests show that DSpace and PubMed may retrieve irrelevant or partially relevant documents.

3. Principles of Efficient Computation

In this section, I focus on principles of efficient computation that can be imposed on Information Retrieval systems in order to narrow down the search space and improve the retrieval of relevant scientific documents. The first principle is to process the functional hierarchical structure of linguistic expressions. The second principle is to process the asymmetric relations between the constituents of linguistic expressions. The processing of functional projections and asymmetric relations contribute to the computational efficiency of the Language Faculty.* I discuss these in turns in the following paragraphs.

* According to Chomsky (2005, 2008, 2013), principles of efficient computation are external to the Language Faculty. Conditions on the derivations and the interface conditions contribute to eliminate complexity, that is, choice points, in the course of the derivation. Interface conditions further reduce the complexity for CI and SM interpretations.

3.1 Functional projections

Efficient information systems must rely on the properties of the hierarchical functional structure of linguistic expressions. In languages, such as English, a sequence of substantive words cannot be interpreted without being part of a functional structure headed by functional elements, such as prepositions and determiners. Consider the examples in (13).

(13) a. Ticlopidine is a potent inhibitor for CYP2C19.
b. Ticlopidine potent inhibitor CYP2C19.

While the proposition in (13a) is straightforwardly interpretable, this is not the case for the expression in (13b), where functional projections are lacking. On one hand, the expression in (13a) can be interpreted by the application of the semantic rule of Functional Application to the constituents of the syntactic structure in (13a). On the other hand, the interpretation of the expression in (13b) is undetermined. Since its functional projections are lacking, it cannot be interpreted as a proposition. Further, (13b) is a set of words functionally unrelated to one another.

Functional structure is necessary for the form and the interpretation of linguistic expressions. The position of functional categories in syntactic structures determines their scope relations. In (14a), the preposition *for* scopes over the DP [$_{DP}$ CYP2C19]. It does not scope over the DP [$_{DP}$ Ticlopidine]. The scope of functional categories contributes to the interpretation of propositions. The propositions in (14a and 14b) have different truth-values. Either (14a) or (14b) is true. They cannot both be interpreted as true propositions in the same world of interpretation.

(14) a. [$_{TP}$ [$_{DP}$ Ticlopidine] T [[$_{VP}$ is [$_{DP}$ a potent inhibitor] [$_{PP}$ for [$_{DP}$ CYP2C19]]]]]
b. [$_{TP}$ [$_{DP}$ CYP2C19] T [[$_{VP}$ is [$_{DP}$ a potent inhibitor] [$_{PP}$ for [$_{DP}$ Ticlopidine]]]]]

The hierarchical position of functional categories plays a role in the semantic interpretation of the linguistic expressions they are part. For example, negation is part of the sentences in (15a) and in (15b). However, the semantic interpretation of these expressions differs as can be seen in (16), where in (16a) the negative element does not scope over the quantifier ‘many’, whereas it is the quantifier that scopes over negation in (16b). It is possible to add the expression ‘just a few’ to (16a) but not to (16b).

(15) a. Not many drugs were tested before the experiment.
b. Many drugs were not tested before the experiment.

(16) a. Not many drugs were tested before the experiment, just a few.
b. Many drugs were not tested before the experiment, just a few.

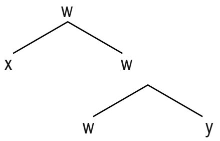
The processing of functional words, such as determiners, quantifiers, negation, and prepositions is crucial for the interpretation of linguistic expressions. Determiners and quantifiers provide generic/specific reference, negation directly affects the truth-value of the propositions, and prepositions identify spatio-temporal semantic relations. Thus, the functional hierarchical structure and the occurrence of specific functional elements must be processed by information systems in order to meet the interface legibility conditions.

3.2 Asymmetrical relations

Linguistic expressions cannot be thought of in terms of unordered sets of words. Given the operations and principles of grammar, linguistic expressions are structured sets of elements, where asymmetrical relations hold. This follows from the *Asymmetry Hypothesis*, according to which asymmetry is a core relation of the Language Faculty (Di Sciullo 2005a). Precedence, dominance, and asymmetrical c-command, defined on binary branching hierarchical structures, are the basic asymmetrical relations for the analysis of linguistic expressions, see (17) for definitions of 'c-command' and 'asymmetrical c-command'. They are central to the form and the interpretation of linguistic expressions. In a model where syntax is the core engine of computation, precedence relations are relevant at the SM interface, while dominance and c-command relations are relevant at the semantic CI interface. Thus in (18), x precedes w and y, x asymmetrically c-commands y: however, y does not asymmetrically c-command x.

(17) a. C-command: X c-commands Y if 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. (Kayne 1994).

(18)



Asymmetry is hard-wired in natural language. With respect to the form of linguistic expressions, asymmetric relations hold for selection and displacement. If a category X selects a category y, the inverse relation does not hold. Thus, a verb selects its DP complement and the inverse relation does not hold. The displacement of constituents is part of the derivation of linguistic expressions. However, as it is the case for selection, syntactic movement is asymmetrical in the sense that if a category X moves to a position Y, the inverse movement does not hold. Selection and displacement are subject to structure dependent principles, such as the c-command relation. A selected constituent is merged as a sister to the selector (c-command) while a displaced constituent can only be remerged to a higher asymmetrical c-command position, as the following structures illustrate.

(19) a.[[Ticlopidine] [is [~~Ticlopidine~~] a potent inhibitor for CYP2C19]]]]
←
b.[[A potent inhibitor for CYP2C19] [is Ticlopidine [~~A potent inhibitor for CYP2C19~~]]]
←

(20) a. They discovered [[genes] of [[genes] this type]]
←
b. They discovered [[this type] of [gene [this type]]]
←

Asymmetrical relations are central relations in the expressions derived by the operations of the Language Faculty. These cannot be thought of in terms of strings of characters or bags of words; linguistic expressions are structured sets of elements. Given the central role of asymmetry in grammar, the following legibility condition must hold at the interface of the grammar and the performance systems, CI and SM.

(21) Interface Legibility Condition:
Asymmetric relations are optimally legible/interpretable by the performance systems.

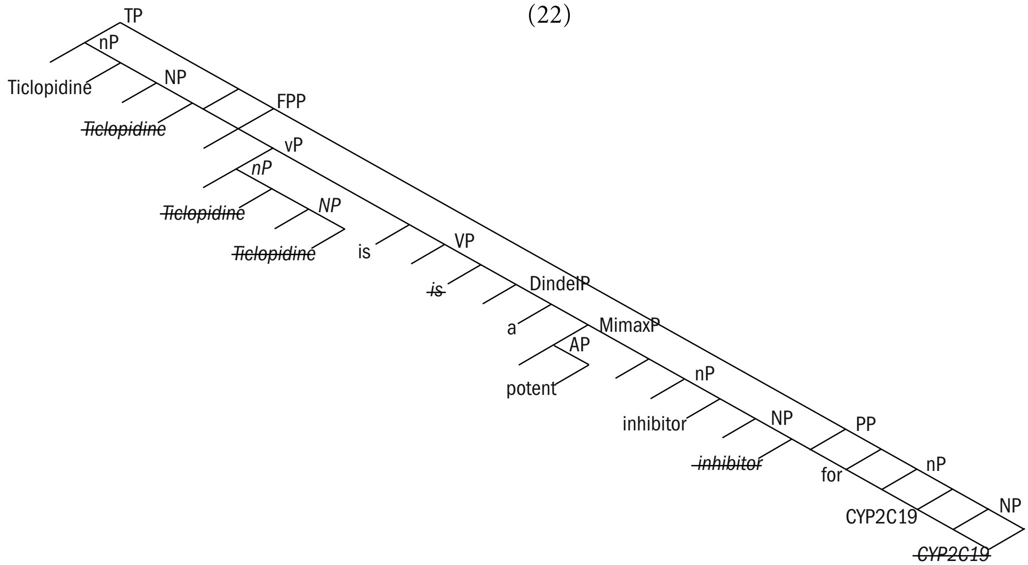
According to the Asymmetry Theory (Di Sciullo 2005a), grammar is the best solution to the interface legibility conditions in the sense that the asymmetry of linguistic expressions matches with the asymmetry of the external systems. Thus, the syntactic operation merge is an asymmetric operation that matches with the asymmetry of the semantic operation of Functional Application.

3.3. Parsing functional projections and asymmetrical relations

The processing of the hierarchical functional structure and the asymmetric relations between the constituents of linguistic expressions are the necessary bases for efficient natural language processing. They are also necessary for information systems, such as Digital Libraries. The LAD parser (Di Sciullo 2010a; Di Sciullo 2012, 2013) ensures the recovery of the functional projections and the asymmetric relations of linguistic expressions.**

The architecture of the parser is such that it limits the search space and the computational actions, while it incrementally recovers the asymmetric structures from the input. The parser assembles phrase structure from left to right. It computes asymmetric relations as it incrementally processes the input one word at a time. Each time a word is introduced, it extends the analysis produced at that point. The model includes mechanisms to implement efficient parsing, without backtracking or unnecessary search in the derivations. The parser efficiently interprets the grammar by restricting the operations of the grammar to apply in local domains. The trace in (22) is an example of the recovery of asymmetrical relations by the LAD parser.

(22)



The parse tree in (22) illustrates that Ticlopidine is the logical subject of the sentence, as it is generated within the vP, before being displaced (remerged) within the TP. This is not the case for CYP2C19, which is an adjunct headed by a functional head, *for*, and is generated outside of the vP. In the parse tree in (12), the logical subject asymmetrically c-commands the adjunct and not conversely. The parser recovers the asymmetry, notwithstanding the fact that

** Computational implementations of asymmetric relations are already available. The asymmetric c-command relation is part of Marcus's parser (Marcus 1980), as well as in Government and Binding computational implementations (Berwick and Weinberg 1984; Berwick 1985; Berwick *et al.* 1991; Fong 1991, 2005, a.o.), and in recent works on asymmetry and minimalism (Di Sciullo 1999, 2000, 2005b, 2013; Di Sciullo and Fong 2005; Harkema 2005; Stabler 2011, 2013, a.o.). A computational model based on the recovery of asymmetric relations leads to a new paradigm in natural language processing.

the logical subject and the adjunct are nominal constituents.

Summarizing, the form of linguistic expressions include functional projections where asymmetrical relations hold. The processing of functional projections and asymmetrical relations are constraints that can be imposed to information systems, such as Digital Libraries, to narrow down the search space and improve the retrieval of relevant documents.

4. Convergences between Biolinguistics and Information Technologies

As language is an object of the natural world, and that asymmetrical relations are part of biology, (see Di Sciullo *et al.* 2010b), it does not come as a surprise that asymmetry also affects the generative procedure of the Language Faculty.

Biolinguistics has implications for natural language technologies, including information retrieval and Digital Libraries. The processing of the asymmetric properties of linguistic expressions enables any area where human users can benefit by communicating with their computers in a natural way.

Asymmetric relations, couched in terms of asymmetric c-command relations between the constituents of linguistic expressions, must be recovered in order to determine the set of documents providing relevant answers to the request of information formulated by the query. Information processing, oriented by the recovery of asymmetric relations contributes to the development of efficient information systems, since it relies on the biological properties of the human Language Faculty.

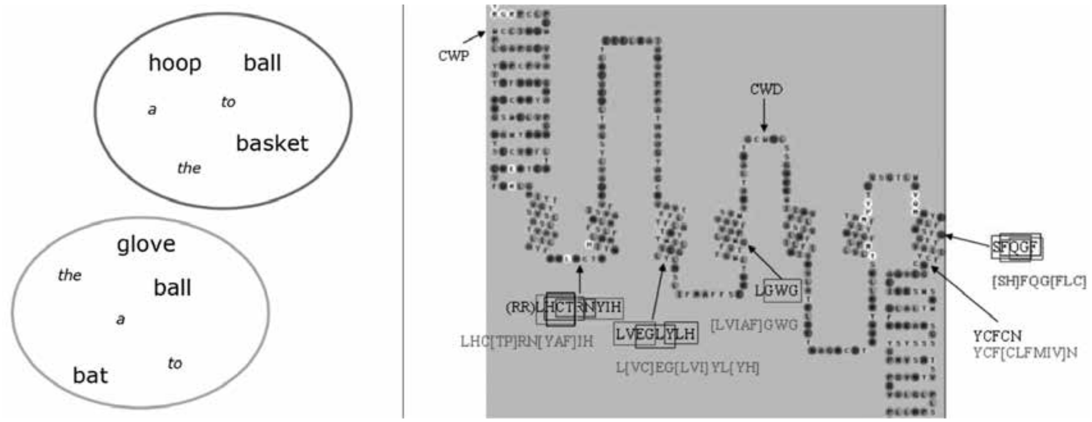
The biolinguistic approach to information systems brings about convergences between linguistics, biology, and information technologies. Given the properties of the Language Faculty and the central complexity-reducing role of asymmetry in the processing of linguistic expressions, efficient information systems should rely on the recovery of the asymmetrical relations of linguistic expressions.

However, the connection between linguistics, biology, and information technologies is, more often than not, based on current word-based practice in information processing, rather than on the asymmetrical properties of the narrow Language Faculty. The following example is an illustration of such a misleading connection. In Cheng *et al.*(2004), an analogy is drawn between feature selection in language and feature selection in biology. The analogy is based on the assumption that substantive words (such as *ball*, *glove*, *bat, basket*) differ fundamentally from functional words (such as *a, to, the*). Some substantive words (e.g., *ball*) identify the relatedness of documents, while other substantive words (e.g., *glove, bat*, and *basket*) identify their differences. Functional words (e.g., *a, to, the*) are irrelevant for distinguishing the documents from each other and from unrelated ones.

According to Cheng *et al.*(2004), word equivalents used in protein sequence language are short stretches of amino acids. Only some amino acid positions are useful in distinguishing different subtypes of a protein-coupled receptor, while the helices (center) are common to all protein-coupled receptors. The other areas cannot be distinguished from any other protein. See (23) which illustrates this further.

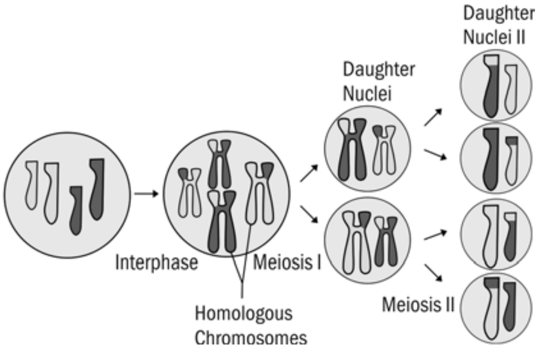
The bag-of-word approach to information processing does not take into account the asymmetry of linguistic relations. Consequently, the relations between the substantive words and other substantive words mediated by the functional elements are lost, and with them the semantic content conveyed by the expressions they are part of. The performance of digital libraries incorporating information retrieval systems, based on the bag-of-word technique and the keyword search is not optimal, as we illustrated earlier in the article. The processing of functional projections instead, will enhance their performance.

(23)



Promising convergences between linguistics, biology, and information technologies are based on the properties of the operations of the language faculty, that is binarity and recursion, as well as on principles of efficient computation such as asymmetry and on the biological basis of these properties, which can be found, for example in the properties of cell replication, where binarity, recursion, hierarchy, and asymmetry are observed, see (24), which is a representation of cell replication.

(24)



Formal properties of relations, such as symmetry and asymmetry, are used to describe the dynamics of morphogenesis in biology (Montell 2008), and to formulate laws of physics. Information processing systems oriented by the recovery of asymmetric relations are likely to be more efficient, since they rely on the principles of efficient computation restricting the computational procedure of the Language Faculty.

5. Summary

In this article, I related Biolinguistics, Natural Language Processing, and Digital Libraries of the future. Linguistic expressions are not strings of characters or bags-of-words; their constituents are organized in a hierarchical structure where asymmetric relations hold. The performance of information systems based on FSG or the bag-of-word technique is not optimal because these systems do not rely on the core properties of the computational procedure of the Language Faculty and the principles of efficient computation. Current 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 processing of functional projections and natural language asymmetric relations, instead of the keywords, is a step forward in the optimization of these systems.

Acknowledgements

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, and by funding from the FQRSC on Dynamic Interfaces, grant number 137253 awarded to Professor Anna Maria Di Sciullo, of the Département de linguistique at the Université du Québec à Montréal. Details available at www.interfaceasymmetry.uqam.ca, www.biolinguistics.uqam.ca

Endnotes

1) The large majority of search engines combine Boolean procedures with another method, including the ones listed below, 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.

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

See Frakes and Baeza-Yates 1992; Strzalkowski 1999; Baeza-Yates and Ribeiro-Neto 1999; Manning, Raghavan and Schutze 2008, a.o., for further discussion.

2) For example, the expression *cba* is derived by the recursive application of Merge to pairs of elements in the Numeration in (ii), as shown in the derivation in (iii).

(i) Merge (a,b) : {a, b}
(ii) Numeration : {a, b, c}
(iii) Derivation:
1. Merge (b, a) : {b, a}
2. Merge (c, {b,a}) : {c, {b, a}}

The principle of prominence determines which of the two merged objects projects its label. The assignment of a label to the derived constituents is subject to the proper subset relation, which falls into the class of the principles of efficient computation. Given the proper subset requirement on the selection of the syntactic objects undergoing Merge, the order of application of this operation is derived without stipulation. See Di Sciullo (2005a) and Di Sciullo and Isac (2008) for discussion.

3) One important contribution to our knowledge of natural languages and the generative procedure that derives their properties goes back to Chomsky's (1957) Theory of formal grammars. According to this theory, formal grammars are ranked according to their generative capacity, finite state grammars being the lowest: (recursively enumerable (context sensitive (context free (finite state grammar)))). In Chomsky's hierarchy, the grammars are associated to automata executing the rules of the grammars. Thus, the automata corresponding to a finite state grammar (FSG), the finite state automata, begins in an initial state, runs through a sequence of states (producing a word or string or words with each transition), and ends in a final state.

A FSG defines a language: the set of sentences that it can produce. FSG can simulate recursion if the recursive node is at the left or at the right edge of a rule. Chomsky (1957) showed that there are strings in English consisting of sentences (hereafter S) related by discontinuous conjunctions such as either or, as in either S or S, that are isomorphic to strings that a FSG cannot produce, in particular (infinite) anbn strings. However, only two rules are needed to produce such strings by a grammar that has the generative power of a context-free grammar (CFG). Moreover, only one rule is necessary within the Minimalist framework to generate such strings.

References

Ahler C B, Fiszman M, Demner-Fushman D, Francois-Lang M, and Rindflesch T C. 2007. **Extracting semantic predications from Medline citations for pharmacogenomics**. *Pacific Symposium on Biocomputing* **12**: 209–220.

Baeza-Yates R and Ribeiro-Neto B. 1999. *Modern Information Retrieval.* Boston, MA: Addison Wesley.

Berwick R. 1985. *The Acquisition of Syntactic Knowledge.* Cambridge, MA: MIT Press.

Berwick R, Abney S, and Tenny C. (eds). 1991. **Principle-based parsing: Computation and psycholinguistics**. *Studies in Linguistics and Philosophy*. Dordrecht: Kluwer.

Berwick R and Weinberg A. 1984. *The Grammatical Basis of Linguistic Performance.* Cambridge, MA: MIT Press.

Buettcher S, Clarke C L A, and Cormack G V. 2010. *Information Retrieval: Implementing and Evaluating Search Engines.* Cambridge, MA: MIT Press.

Candela L D, Castelli N Ferro, Ioannidis Y, Koutrika G, Meghini C, Pagano P, Ross S, Soergel D, Agosti M, Dobreva M, Katifori V, and Schuldt H. 2008. The DELOS Digital Library Reference Model–Foundations for Digital Libraries. Version 0.98, February 2008.

Candela L, Athanasopoulos G, Castelli D, El Raheb K, Innocenti P, Ioannidis Y, Katifori A, Nika A, Vullo G, and Ross S. 2011. The Digital Library Reference Model. DL.org Project Deliverable.

Cheng B Y M, Carbonell J, and Seetharaman J K. 2004. **Biolinguistics: The Use of Analogies for Interdisciplinary Research**. Available at <http://www.andrew.cmu.edu/course/60-427/aisd/biolinguistics.pdf>

Chomsky N. 1957. *Syntactic Structures.* The Hague: Mouton.

Chomsky N. 1995. *The Minimalist Program.* Cambridge, MA: The MIT Press.

Chomsky N. 2001. ***Derivation by Phase.*** In *Ken Hale: A Life in Language*, pp. 1–52 edited by M Kenstowicz. Cambridge, MA: MIT Press.

Chomsky N. 2005. **Three factors in language design**. *Linguistic Inquiry* **36**(1): 1–22.

Chomsky N. 2008. **The Biolinguistic Program: Where Does it Stand Today?** Ms. MIT. Available at <http://clt.blcu.edu.cn/resource/[[1].Chomsky_2008_bio.todayGood.gd.pdf>

Chomsky N. 2013. **Problems of projection**. *Lingua* **130**: 33–49.

Chomsky N. 2014. *The Minimalist Program, 20th Anniversary Edition.* Cambridge, MA: MIT Press.

Di Sciullo A M. 1999. **An Integrated Competence-Performance Model, A Prototype for Morpho-Conceptual Parsing and Consequences for Information Processing.** *Proceedings of VEXTAL*, 369-379. Università Ca'Foscari, Venezia.

Di Sciullo A M. 2000. **Parsing Asymmetries**. In *Natural Language Processing. Lecture Notes in Computer Science*, pp. 24–39 edited by D N Christodoulakis. Springer Computer Science Press.

Di Sciullo A M. 2005a. *Asymmetry in Morphology*. Cambridge, MA: MIT Press.

Di Sciullo A M. 2005b. **Domains of Argument Structure Asymmetries**. WMSCI-2005. The 9th World Multi-Conference on Systemics, Cybernetics and Informatics, pp. 316–320.

Di Sciullo A M. 2011. **A Biolinguistic Approach to Variation**. In *The Biolinguistic Entreprise: New Perspectives on the Evolution and Nature of the Human Language Faculty*, pp. 305–328 edited by A M Di Sciullo and C Boeckx. Oxford: Oxford University Press.

Di Sciullo A M. 2012. **Asymmetric Agreement in Pronominal Anaphora**. In *Frontiers in Artificial Intelligence and Applications*, pp. 395–410 edited by H Fujita and R Revetria 246. IOS Press, Amsterdam.

Di Sciullo A M. 2013. **A Reason to Optimize Information Processing with a Core Property of Natural Language**. In *Proceedings of the 9th SoMeT_10*, pp. 437–456.

Di Sciullo A M. 2014. **Minimalism and I-Morphology**. In *Minimalism and Beyond: Radicalizing the Interfaces*, pp. 267–286 edited by P Kosta, S Franks and T Radeva-Bork. Amsterdam: John Benjamins.

Di Sciullo A M and Boeckx C (eds). 2011. *The Biolinguistic Enterprise: New Perspectives on the Evolution and Nature of the Human Language Faculty*. Oxford: Oxford University Press.

Di Sciullo A M and Fong S. 2005. **Morpho-Syntax Parsing**. In *UG and External Systems. Language, Brain and Computation*, pp. 247–268 edited by A M Di Sciullo. Amsterdam: John Benjamins.

Di Sciullo A M and Isac D. 2008. **Asymmetric merge**. *Biolinguistics* **2**: 260–290.

Di Sciullo A M, Gabrini, P, Batori C, and Somesfalean S. 2010a. **Asymmetry, the Grammar, and the Parser**. In *Linguistica e Modelli Tecnologici di Ricerca*, pp. 477–494 edited by G Ferrari, R Benatti, and M Mosca. Roma: Bulzoni.

Di Sciullo A M, Piattelli-Palmarini M, Wexler K, Berwick R C, Boeckx C, Jenkins L, Uriagereka J, Stromsworld K, Cheng L, Harley H, Wedel A, McGilvray J, van Gelderen E, and Bever G T. 2010b. **The biological nature of human language**. *Biolinguistics* **4**: 4–34.

Epstein S and Seely T D (eds). 2002. *Derivation and Explanation in the Minimalist Program*. Malden: Blackwell.

Fitch W T and Hauser M. 2004. **Computational constraints on syntactic processing in a nonhuman primate**. *Science* **303**: 377–380.

Fong S. 1991. **Computational Properties of Principle-based Grammatical Theories**. Ph.D. Dissertation. Artificial Intelligence Laboratory. MIT.

Fong S. 2005. **Computation with Probes and Goals**. In *UG and External Systems. Language, Brain and Computation*, pp. 311–334 edited by A M Di Sciullo. Amsterdam: John Benjamins.

Frakes W B and Baeza-Yates R. 1992. *Information Retrieval*. Upper Saddle River, NJ: Prentice Hall.

Friederici A D. 2009. **The Brain Differentiates Hierarchical and Probabilistic Grammars**. In *Of Minds and Language*: *A Dialogue with Naom Chomsky in the Basque Country* M. Piattelli-*Palmarini*, pp. 184–194 edited by J Uriagereka and P Salaburu. New York: Oxford University Press.

Harkema H. 2005. **Minimalist Languages and the Correct Prefix Property**. In *UG and External Systems. Language, Brain and Computation*, pp. 289–310 edited by A M Di Sciullo. Amsterdam: John Benjamins.

Hauser M D, Chomsky N and Fitch W T. 2002. **The faculty of language: What is it, who has it, and how did it evolve?** *Science* **298**:1569–1579.

Heim I and Kratzer A. 1998. *Semantics in Generative Grammar***.** Malden: Blackwell.

Jenkins L. 2000. *Biolinguistics***.** Cambridge, MA: MIT Press.

Jenkins L. 2004. *Variation and Universals in Biolinguistics***.** Amsterdam: Elsevier.

Karamanis N, Lewin I, Sealy R, Drysdaley R, and Briscoe E. 2007. **Integrating natural language processing with flybase curation**. *Pacific Symposium on Biocomputing* **12**: 245–256.

Kayne R. 1994. *The Antisymmetry of Syntax***.** Cambridge, MA: MIT Press.

Kayne R. 2010. **Why Are There No Directionality Parameters?** In *Proceedings of the 28th West Coast Conference on Formal Linguistics*, pp. 1–23 edited by M Byram Washburn, K McKinney-Bock, E Varis, A Sawyer, and B Tomaszewicz. Somerville, MA: Cascadilla Proceedings Project.

Lenneberg E H. 1967. *Biological Foundations of Language***.** New York: Wiley.

Manning C, Raghavan P, and Schutze H. 2008. *Introduction to Information Retrieval***.** New York, NY: Cambridge University Press.

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

Montell D J. 2008. **Morphogenetic cell movements: Diversity from modular mechanical properties**. *Science* **322**:1502–1505.

Moro A. 2000. *Dynamic Antisymmetry***.** Cambridge, Mass.: MIT Press.

Moro A. 2010. *The Boundaries of Babel: The Brain and the Enigma of Impossible Languages***.** Cambridge, MA: MIT Press.

Murphy R A, Mondragon E, and Murphy V. 2008. **Rule learning by rats**. *Science* **319**:1849–1851.

Neveol A, Shooshan S E, Humphrey S M, Rindflesh T C, and Aronson A R. 2007. **Multiple approaches to fine-grained indexing of the biomedical literature**. *Pacific Symposium on Biocomputing* **12**: 292–303.

Petrides M, and Pandya D N. 1994. **Comparative Architectonic Analysis of the Human and the Macaque Frontal Cortex**. In *Handbook of Neuropsychology*, Vol. 9, pp. 17–58 edited by F Boller and J Grafman. Amsterdam: Elsevier.

Schatz B R. 1997. **Information retrieval in digital libraries: Bringing search to the Net**. *Science* **275**: 327–334.

Stabler E. 2011. **Computational Perspectives on Minimalism.** In *Oxford Handbook of Linguistic Minimalism*, revised version, pp. 617–642 edited by C Boeckx.

Stabler E. 2013. **Two models of minimalist, incremental syntactic analysis** (revised version). *Topics in Cognitive Science* **5**(3): 611–633.

Strzalkowski T. (ed.). 1999. *Natural Language Information Retrieval*. Dordrecht: Kluwer.

Yu H and Kaufman K. 2007. **A cognitive evaluation of four online search engines for answering definitional questions posed by physicians**. *Pacific Symposium on Biocomputing* **12**: 328–339.