An Integrated Competence-Performance Model, A Prototype for Morpho-Conceptual Parsing and Consequences for Information Processing

Anna-Maria Di Sciullo
Université du Québec à Montréal

Introduction[1]

We present the basis of an integrated model of language knowledge and language use relying on asymmetry. We consider the properties of morphological structures and argue that asymmetry contributes to the analysis of their form and interpretation. We implement our model in a morphological parser and compare the properties of this parser with another morphological analyzer.

As existing information processing systems are beginning to integrate articulated lexical knowledge (Copestake & Briscoe, 1996; Paziena 1997; Boguraev & Pustejovsky, 1996, and related works) we expect that a refinement of their architecture with asymmetry-based parsing will increase their performance.

The organization of this paper is the following. The first part presents our integrated model. The second part identifies the main features of a prototype for morphological analysis incorporating the asymmetry-based grammar. The last section considers the consequences of the inclusion of the parser in information processing systems.

1. An integrated Model

We posit the hypothesis in (1) on the interaction of the grammar with the performance systems, and define the logical relation of asymmetry in (2).

(1) **Integrated Asymmetry Hypothesis**

a. Universal Grammar is designed to optimally analyze linguistic expressions in terms of asymmetrical relations.

b. Universal Parser is designed to optimally recover natural language asymmetries.

(2) **Asymmetrical Relation**

x and y are in asymmetrical relation, iff there is a formal relation r such that r(x, y) does not imply r(y, x).

The following identifies the specificity of our proposal with respect to the theory of grammar as well as with respect to parsing theory.[2]

With respect to the theory of grammar, our proposal is compatible with Kayne's (1994) anti-symmetry framework as well as with Chomsky's (1998) Minimalist Program. In both cases, asymmetrical relations are part of the operations and the conditions determining the linear order of terminals,[3] as well as the restrictions on the composition and attraction of non terminals.[4]

The contribution of our proposal to these lines of research lies on the centrality of (2) in the grammar. Firstly, phenomena previously accounted for in terms of symmetrical c-command, or in terms of bare sisterhood, are reanalyzed in terms of asymmetry. Secondly, our proposal extend to areas that have not yet been analyzed in terms of asymmetry, including

---

[1] This work is supported in part by funding from the Social Sciences and Humanities Research Council of Canada to the Asymmetry Project, grant number 214-97-0016.

[2] Singular grammars instantiate specific parameters of Universal Grammar (UG) (Chomsky, 1981, 1998, and related works) and parsing is the processing of UG principles and parameters (Berwick, 1991; Fong, 1991; Door, 1991, Kashket, 1991 and related works).

[3] According to the Linear Correspondence Axiom (Kayne, 1994), the linear order of terminals is a function of the asymmetrical c-command relation of the non-terminals.

(i) X c-commands Y iff X and Y are categories and X excludes Y and every category that dominates X dominates Y (Kayne 1994:16).

See also Chomsky (1998), Epstein (1995), Reuland (1997), Robert & Vijashankar (1995) for different definitions of c-command.

[4] Thus, for example, in Chomsky (1998), the asymmetrical operation PAIR-MERGE derives adjunction structures, while SET-MERGE, further restricted by the selectional properties of heads, derives head-complement structures.

lexical representations, and morphological configurations, as discussed below.

With respect to the theory of parsing, the specificity of our proposal lies in the close connection between the grammar and the parser. The parser recovers the structures of linguistic expressions on the basis of the analysis of local asymmetrical relations. The structures include both formal (categorial) and semantic (argument, aspect and concept) features, allowing for a full recovery at each step of the parse.

The specificity of our integrated model lies on the hypothesis that while the grammar derives complex feature structures, no feature can be interpreted optimally by the performance systems if not presented in canonical asymmetrical formats. Thus, asymmetry is also central to the interface condition ensuring feature visibility/ interpretation by the performance systems (Conceptual-Intentional (C-I) and Acoustic-Perceptual (A-P)).

1.1 Model

1.1.1 Grammar

Our proposal restricts the Minimalist architecture of the grammar to the following.

```
(3)           Morpho        Computational
      Lexicon Syntax            Space
              Phono             /   \
                              LF     PF   (IUA)
                              C-I    A-P
```

The restrictions cover each module of the grammar, viz., the lexicon, the components (morphology, syntax and phonology), the computational space, the interfaces (Logical Form (LF) and Phonetic Form (PF)) as well as the condition on the structural descriptions generated by the grammar, the IUA.

The first restrictive property of our model is that the lexical features are presented in the same format. Both the formal and the semantic features associated with lexical items are encoded in terms of the same set of asymmetrical relations. [5]

The second restrictive property of our model is that, even though autonomous with respect to the sort of features they include, the operations of each component generate asymmetrical relations.The derivation of the different sorts of linguistic expressions, words and phrases, takes place in the same space and is driven by the conceptual necessity to obtain Canonical Target Configurations (CTC) at the interfaces with the performance systems, as proposed in Di Sciullo (1996).[6] These configurations, generally refered to as specifier-head, head-complement and adjunct-head relations, are asymmetrical relations. They satisfy the definition in (2), as, in each case, there is an unidirectional relation that goes from one member of the relation to the other and not conversely. Assuming that the grammar manipulates feature structures, the formal head-dependent asymmetrical relations can be defined as follows:[7]

(4) The features of a complement are properly included in the features of the head it is a complement of .

(5) The features of a specifier are partially excluded from the features of the category it is the specifier of .

(6) The features of an adjunct are excluded from the features of the category it is adjoined to.

Interpreted in terms of inclusion and exclusion relations between feature structures, the formal head-dependent asymmetries are distinct but nevertheless related relations. This makes complements distinct from adjuncts (Chomsky, 1995; Kayne, 1994) while it allows the identification of specifiers to adjuncts (Kayne, 1994).[8]

The formal (categorial) head-dependent asymmetries in (4)-(6) are not isomorphic to the semantic asymmetries in (7)-(9), as the same semantic relation may be supported by different formal head-dependent relations.

---

[5] This differs from the standard assumptions, (Lieber, 1992, among others) where different notations are used to encode different lexical properties, such as sub-categorization frames, argument structures and lexical conceptual structures.

[6] This property of our model makes it distinct from current models of grammar. For example, it differs from a model incorporating Distributed Morphology (Hale and Marantz, 1993), where parts of morphological derivations are done by the rules of other components. Our model also differs from a theory incorporating Anderson's (1992) A-morphous Morphology, where morphological rules may insert/eliminate/reintroduce affix morphemes in the derivations.

[7] Objects are in included in the VP, (4), while subjects and adjuncts are not , (5), (6); subjects are external to the VP but internal to v, (5); adjuncts are external to [v [VP]], (6).

[8] Our model is compatible with Cinque's (1997) system, where syntactic modification, either adverbial or adjectival, is restricted to specifier-head relations in the functional projection of the lexical categories. In our model, adjunct-head relations are not part of the derivation of syntactic expressions, they are limited to the generation of morphological expressions.

(7) An argument satisfies an argument position of the predicate it is an argument of.
(8) An operator binds a variable is its local domain.
(9) A modifier identifies an unspecified feature of the category it is the modifier of.

The relations in (7)-(9) also qualify as asymmetrical relations. Each one instantiates a unidirectional relation that goes from one member of the relation to the other and not conversely.

In our model, conceptual features are visible/interpretable at the conceptual interface only if they are part of CTCs.

Thus, in derivational morphology, the predicate-argument asymmetry is instantiated in the relation between category-changing affixes and roots; the modifier-modified asymmetry is licensed in the relation between scalar affixes and roots; the operator-variable asymmetry is licensed in the relation between superlative affixes and roots. While the semantic asymmetries are canonically supported by the adjunct-head relation under the word-level, they are supported by the complement-head and the specifier-head relations in phrasal syntax. The following examples illustrate this point.

(10) a. Nobody is *able* to drink this.
b.This is not drink*able*.
(11) a. Zev protects Stan *too much*.
b. Zev *over*protects Stan.
(12) a. Homer is *more* happy than Bart.
b. Homer is happi*er* than Bart.

In (10a), the predicate-argument relation is instantiated by the head-complement relation holding between the verb *able* and the VP *to drink this*, and by the adjunct-head relation in (10b), relating the adjectival affix -*able* and the verb *drink*. In (11a), the modifier-modified relation is instantiated by the specifier-head relation relating the adverbial phrase *too much* to the VP *protects Stan* and by the head-adjunct relation in (11b), relating the prefix *over* to the verb *protect*. In (12a), the operator-variable relation is instantiated by the superlative operator *more* in specifier-head relation with the variable it binds, and in (12b) by the superlative suffix *–er* in adjunct-head relation with the adjective *happy*. Thus, formal asymmetrical relations range over syntactic and morphological expressions in specific ways.

Another restrictive feature of our model is that it includes a unique condition on the interpretation of the expressions generated by the grammar.

(13) **Interpretation Under Asymmetry** (IUA)
a. An interpretation is optimally obtained under a unique local asymmetrical relation.
b. A local asymmetrical relation optimally supports a unique interpretation.

Our model is distinct from current models of grammar and related parsing systems. It differs for example from GB theory (Chomsky 1981, 1986) as the integrated asymmetry model includes a unique condition on derivation and representation contrary to the multiple GB sub-principles (X-bar, Case-theory, Government Theory, Binding Theory, Theta-Theory). The IUA subsumes the heterogeneous GB sub-principles, which, in our view, are particular instantiations of the basic asymmetry of the grammar and need not to be stipulated as such.

1.1.2 Parser

We propose an integrated theory of language, where the performance system is designed to use the asymmetrical properties of the linguistic expressions generated by the grammar in an optimal way. In this perspective, let us take the general principles underlying asymmetry-based parsing to be the following :

(14) **Asymmetry-Based Parsing**
a. The parser makes an optimal use of the asymmetrical relations of the grammar.
b. The operations of the parser are controlled by the IUA.
c. The parser provides an incremental analysis of the linguistic expressions.

We are thus proposing a direct implementation of the asymmetry of the grammar in the computational model: the asymmetrical properties of the competence grammar are directly used by the processor. That is, the system integrates the universal asymmetrical relations as well as the grammar specific parametric values. The lexical database includes the idiosyncratic properties of underived items, couched in terms of asymmetrical relations. The actions of the parser are oriented by the identification of local asymmetrical relations at each step of the parse, including categorization, attachments and dependencies between categories. In our model, the computational operations are oriented by natural language asymmetries and not by language independent heuristics. The IUA is used as an overall control devise that ensures the legitimacy of the choices undertaken by the parser.

(15) **Control**
a. Search until an asymmetrical relation is identified.
b. Search for formal and semantic asymmetry at each step of the parse.
c. In case of multiple choice, choose the more local asymmetrical relation rather than a non-local one.

Our parsing model differs from the generate and filter model that implemented GB Theory, as the decisions of the parser at each step of the parse are oriented by the recovering of the asymmetrical relations. We expect that the problems of over-generation and speed associated with the early principled-based parsers to be significantly reduced in a parsing model that incorporates the asymmetry-based grammar.

That the IUA acts as an overall control mechanism is motivated by our understanding of the processing of structural complexity.

Structural complexity arises in situations where more than one asymmetrical relation is available at a given of the parse. The complexity in the processing of multiple dependencies and garden paths can be treated in a unified way, as we will see immediately.

Multiple attachment are cases where a given semantic asymmetrical relation can be supported by more than one formal relation of the same sort. Garden paths are cases where different sorts of asymmetrical relations are available.

Alongside the phrasal cases of multiple attachments, illustrated in (16a,b), structural complexity may also arise under the word level, as exemplified in (16c,d).

(16) a. John saw Bill on the hill.
b. John photographed Bill with a knife.
c. John rewired the house.
d. John reopened the theater.

The structural complexity arising from prefixed verbal constructions, such as (16c), is a consequence of the licensing of more than one local asymmetrical relations, inducing a repetitive/restitutive interpretation or only a repetitive interpretation for the event denoted by the verbal predicate.

Garden paths are worse to process than multiple attachments, as, contrary to the latter, they present a situation where different sorts of asymmetrical relations are possible at a given point. With syntactic garden paths such as (17a), *the horse* is miss-analyzed as the subject of the VP, a specifier-head relation, instead of being analyzed as the head of a restrictive relative clause. The presence of the complementizer *that* in (17b) dismisses the garden path effect. With morphological garden path, such as (17c), the iterative prefix is miss-analyzed as part of a head-complement relation within the verbal projection, instead of being analyzed as an external modifier of the verbal projection. The presence of the intervening internal prefix *en-* in (17d), entering in a head-complement relation within the verbal projection, forces an adjunct-head analysis for the iterative prefix, and dismisses the garden path effect.

(17) a. #The horse went past the barn fell.
b. The horse [that went past the barn fell]
c. #They reforced their positions.
d. They [re[en-forced]] their positions

Garden paths are worse for the processor than multiple attachments because, contrary to the latter, the former presents a situation where more than one sort of asymmetrical relations are virtually possible at a given point of the parse in the grammar of the language under consideration.

The IUA predicts that the optimal interpretation will be the one induced by the more local asymmetrical relation. This makes the correct prediction for multiple PP attachments, as in (16a,b), where the PP is optimally analyzed as a modifier of the VP and not as a modifier of the whole sentence, as well as for the restitutive/repetitive interpretation of the prefixed verbal structures in (16c,d). In the case of garden paths, the IUA also makes the right prediction. The local asymmetrical relation is prefered and leads to miss-analysis.

1.1.3 Summary

In our integrated model, the grammar and the performance systems manipulate linguistic expressions via unique local asymmetrical relations. Morphological as well as syntactic expressions containing either polysemic morphemes or multiple formal asymmetrical relations at a given point are not interpreted optimally.

2. A Prototype for morpho-conceptual parsing

We present the main features of a prototype implementing the asymmetry-based grammar for the analysis of morphological structure. We refer to this prototype a CONCE-MORPHO-PARSE.[9] The prototype is implemented by Christian Thérien in the Asymmetry project at the Université of Québec à Montréal.

[9] CONCE-MORPHO-PARSE is a refinement of MORPHO-PARSE (Di Sciullo, 1997b), which was designed to analyze the categorial and argument-structure properties of complex words. While MORPHO-PARSE implemented the morphological theory of Di Sciullo and Williams (1987), CONCE-MORPHO-PARSE implements the theory of asymmetry of Di Sciullo (1997a, 1998, 1999). It analyzes word-internal formal and semantic asymmetries.

The morphological parsing is performed by a Unification Grammar and a LR(1) control structure. The prototype builds morphological trees for words (W) identifying the morphemes (Head (H), External Prefix (EP), Internal Prefix (IP), Suffix (S), Root (R)) It recovers the asymmetrical relations, (the adjunct-head (AH), the head-complement (HC) and the specifier-head (SH) relations) the morphemes are a part of.

A unification-based chart parser, the general properties of which are described in Shieber (1986), provides the parse trees with categorial and conceptual feature structures. Unification is useful in implementing an asymmetry-based grammar, as asymmetry holds primarily for pairs of feature structures. Thus, feature unification under head-complement asymmetry is possible only when the features of the complement are properly included in the features of the head; feature unification under adjunct-head relation is possible only if the features of the adjunct are not included in the features of the head; feature identification under specifier-head asymmetry is possible only if the features of the head are partially excluded from the feature structure of its head. The LR(1) grammar controls Unification and implements the IUA Condition, as the operations of the grammar apply only if the relevant symmetrical relation between two feature structures is met. The prototype builds trees on the basis of the recognition of head-complement (HC), adjunct-head (AH) and specifier-head (SH) relations, as only these asymmetrical relations are accepted by the parser.[10]

(18) **Part of the LALR(1) grammar**

a. W -> H
b. W -> HC
c. W -> AH
d. HC -> HC S
e. HC -> R S
f. HC -> PI R
g. AH -> HC S
h. AH -> PE AH
i. AH -> PE HC
j. AH -> P R
k. H -> R

We present here the analysis of the left edge of words with CONCE-MORPHO-PARSE, incorporating the morpho-conceptual analysis of prefixes of Di Sciullo (1977a) in the asymmetry-based grammar.

[10] The singular rules of the LR(K) grammar can be subsumed under a more general structure building operation oriented by asymmetry. We will not discuss the latter implementation of the asymmetry-based parsing model here.

The parser analyses a string of morphemes from left to right and assigns, for every word (W), a parse tree which correctly differentiates the external prefixes, basically modifiers and operators, from internal prefixes, basically predicates. This makes the correct predictions with respect to the argument structure properties of prefixed words, in particular, that external prefixes do not change argument structure while internal prefixes do, as well as it accounts for their aspectual and quantificational properties. This is not the case for current morphological analyzers, where all prefixes are treated on a par, as we will illustrate below.

In our model, the formal asymmetrical relations are paired with the semantic asymmetrical relations predicate-argument, modifier-modified and operator-variable relations. While category-changing suffixes are part of predicate-argument asymmetry, derivational prefixes also participate in the modifier-modified asymmetry as well as in the operator-variable asymmetry. CONCE-MORPHO-PARSE recovers these asymmetries given the Unification grammar, the LR(1) control structure and the lexicon, an extract of which is given below.

(19) **Extract of the lexicon**

re-: Rel_F (x), MOD (e), AGAIN (s)
un-: Rel_F (x), MOD (e), INVERSE (t)
de-: Rel (x), PRED (r), OUT OF (loc)
en-: Rel (x), PRED (r), IN (loc) or IN (state)
a-: Rel (x), PRED (r), TO (loc) or IN (state)

CONCE-MORPHO-PARSE recovers the predicate-argument, modifier-modified as well as the operator-variable asymmetries. The first relation is licensed under argument saturation, the second relation is licensed when there is an identification relation, the third relation is licensed when an operator binds a variable.

The parse trees derived by CONCE-MORPHO-PARSE present a fine-grained analysis of complex words, as illustrated here with the parse tree in (20) for the prefixed de-adjectival verb *re-enlarge*. This is not the case for other morphological analyzers, for example PC-KIMMO 2 with Englex 2 (Karttunen, 1983; Antworth, 1990, and related works), which would assign the parse tree in (21) to this expression.

```
(20)            W
                 !
                AH
           PE   --!--    AH
          Re+       HC-- !-- S
        [RelF]       !        [V]
     [AGAIN]       !        [BECOME]
                 PI - -! -- R
                [Rel]     [state]
                [IN]        large
                 en+
```

```
(21)            Word
                 !
                Stem
      Prefix-----!-----Stem
        re+      Prefix---!---Stem
     ORD5            !            !
                    en+          Root
                REV2+          large
```

Categorially, CONCE-MORPHOPARSE associates an adverbial feature (Rel_F) to the prefix *re-*, which is the adjunct in the asymmetrical adjunct-head relation formed by the prefix and the rest of the structure. However, this asymmetrical relation cannot be recovered at this point of the parse, as there is no such relation between the iterative prefix and the directional prefix *en-*, which is the second element analyzed by the parser. Rule (18h) of the control structure is thus on hold. As there is no asymmetry between the first and the second prefix, they do not form a constituent. There is however an asymmetry between the directional prefix *en-* and the root adjective *large*. It is a head-complement relation, given the lexical specification of the prefix *en-*, given in (19) above, and the adjectival root *large*. The Unification Grammar, controlled by rule (18f), correctly parse the head-complement relation. This relation, encoding a more local asymmetrical relation, is parsed before the adjunct-head relation, controlled by rule (18g). This ensures the connection of the already parsed head-complement structure to the covert verbal head of the structure. Finally, the Unification grammar controlled by rule (18h) attaches the external prefix to the lastly parsed relation.

It is crucial to note here that the category of the covert head of the structure [V] as well as its conceptual feature [BECOME] are derived by the parser, as the prefix *re-* locally identifies the verbal event (e) category it is adjoined to, as specified in its lexical entry in (19). In the case at hand, the event denoted by the predicate is inchoative [BECOME], given that the verbal projection includes an adjective. Inchoative verbs are change of state predicates and states can be categorially realized as adjectives.

The configurational as well as conceptual distinctions between on the one hand complementation-predication and on the other hand adjunction-modification, are not expressed by current morphological parsers, such as Englex and PC-KIMMO for example. Furthermore, there is no way for a parser that does not license phonetically null categories to assign the correct categorial structure to deadjectival (and denominal) verbs in English without loosing the generalization expressed by the Relativized Head Rule according to which the categorial head of the word is the right most member of that word marked for categorial features. CONCE-MORPHO-PARSE recovers covert structures on the basis of formal and semantic asymmetries, given independently needed conditions such as the Relativized Head (Di Sciullo and Williams, 1987) and the Adjunct Identification Condition (Di Sciullo, 1997).

Summing up, our prototype assigns the correct categorial and semantic structure to complex words on the basis of the recovering of formal and semantic asymmetries. Moreover, it is able to restore covert formal and semantic feature structures on the basis of decisions taken locally.

3. Information processing

In this section, we present a view of information processing (IP) that relies on asymmetry-based natural language processing. Our line of reasoning applies to IP systems in general, including information extraction and information retrieval systems, even though each one differs with respect to internal architecture, sort of inquiry, sort of results and evaluation measures. We will not discuss of these differences here.

3.1 Architecture

Amongst IP architectures incorporating linguistic modules, we find the following for Information Extraction systems from Basili and Pazienza (1997: 60), which includes grammatical as well as lexical modules.

(22) **IE architecture**

```
Source doc. → Ling. preprocessor   ← Background lex
|              ←                                    |
   Doc. Annotations ----------------------------    |
          ↑                  ↑                 ↓     ↓
|         |
          |    Doc. Concept Repre ←  Semantic interpreter
|         -----------                        shallow
      Document      |
|      Grammar      |       |
          |         |
|         |         |       ↓
          →    Extraction/Annotation   ←  Domain lexicon
|---------→             Module
                          ↓
                       Template
```

The following diagram presents the modules of a linguistic processor incorporating morpho-conceptual and syntactico-conceptual analysis.

(23) **Linguistic processor**

```
Source documents  →      stemmer   →        Function-
  |                         ↓              Conceptual
                                              lexicon
  |              morpho-conceptual   analyzer

  |                         ↓
               lexico-conceptual  POS tagger    Morpho-
                                               Conceptual
                            ↓                   lexicon
  ↓             syntactico-conceptual analyzer
                            ↓
Document syntactico   ←   syntactico- conceptual
Conceptual representation       super-tager
```

We expect the integration of CONCE-MORPHO-PARSE and other asymmetry-based modules in IP systems to optimize their performance.

3.2 Properties of the system

In the following, we outline the main motivations for a linguistic processor incorporating morpho-conceptual parsing.

While stemming algorithms, such as Porter (1980) and Antworth (1990), do not fully rely on word internal information, in our model, parts of words provide categorial and conceptual features that contribute to the form and content of the information they convey. The processing of these features is necessary for IP in more than one way. Morpho-conceptual features are part of the feature structures of singular words and they constitute the basis upon which decisions about the properties of their local syntactic context can be anticipated. These predictions are not based on statistical calculus, but rather on morpho-conceptual knowledge.

In our model, IP reduces to one case of grammar use by a performance systems. That is, we take the performance systems to be capable to use the formal properties of the grammar to extract or retrieve information from texts. Thus, an IP system is an automated performance system that interprets the linguistic expressions it is exposed to on the basis of the asymmetrical properties of these expressions in order to retrieve or extract specific information. We take these properties to be mainly grammatical in nature and not exclusively statistical.

However, most functioning IP systems are exclusively based on stochastic methods, such as the Boolean search criterion.[11] The use of such methods rely on the view that the representation of the meaning of natural language expressions is independent of the very properties of these expressions. On the other hand, morpho-conceptual systems are natural language dependent. Even though, as it is the case for any system, they constitute a simplification of the object they define, they are similar to their object, in such a way as to allow one to obtain new knowledge about it.

While it is acknowledged that IP systems based on strictly probabilistic methods achieve high levels of performance (Derose, 1988; Marken, 1990, and retated works), it is generally admitted that these systems have now met their limits. Several works indicate that the use of natural language processing methods in IP systems contribute to improve the performance of these systems (Pohlmann & Kraaij 1997; Pustejovsky, Boguraev, Verhagen, Buitelaar, Johnston 1997, as related works).

Different factors contribute to the non-optimality of IP systems operating on strictly probabilistic methods. One factor is that stemming algorithms, such as Porter's (1980) algorithm, do not remove affixes on the basis of the morphological structure of the languages under consideration. Such algorithms cannot be parametrized to apply to languages with non-concatenative morphology, as it is the case for Arabic. Contrastingly, IP systems based on natural language asymmetries may do so.

Another factor that limits the performance of IP systems is that documents are indexed on the basis of unanalyzed lexical items (key words). Such a processing is not optimal because the meaning/information conveyed by texts is

[11] In Information Retrieval systems for example, the indexing of the query consists in its translation into Boolean proper forms, while the indexing of the documents consists in the identification of document profiles by means of the selection of descriptors (significant word) for document profiles using a descriptor dictionary.

not expressed by unanalyzed lexical item. In our model, the information conveyed by lexical items is a function of the asymmetrical relations holding between their parts, as we illustrated above in (19). It is likely that IP algorithms that recover word internal conceptual feature structures will present a closer approximation of semantic information than algorithms exclusively based on stochastic methods.

Our proposal affects the following modules of the linguistic processor: the stemmer, identifying the canonical form of words, the morphological analyzer, identifying word-internal structure, the part-of-speech taggers, identifying the category of lexical items and the post-taggers, identifying the constituent structure of phrases. We expect that IP systems will gain in optimality if they include asymmetrical relations. Such systems will achieve the following:

(24) a. a fine-grained conceptual analysis of word-structure
b.a fine-grained categorial analysis of word-structure
c. disambiguation before part-of-speech tagging
d. greater precision in super-tagging

Assuming that IP systems are series of texts filters, as depicted above in (23). The first filter submits a text to morphological and lexical analysis, assigning tags, including morpho-conceptual feature structures to words. The dictionary plays an important part in tagging. In particular, the lexical entries for derivational affixes and functional categories carry important information about the syntactic and the semantic structure of the linguistic expressions. The initial tagging is performed on the basis of a limited dictionary consisting of function words. The functions words are specified in the dictionary in our features under asymmetry format, as illustrated in (25) for *to*, which is either a preposition (Rel) or a complementizer (Rel_F).

(25) to: Rel (x), PRED (r), TO (Loc)
Rel_F (x), OP (e), INF (e)

A comparison of the text words against the dictionary is performed, a sentence at a time, by a sequential merging process. As a result of the look-up process, any word found in the text will have received one or more tags.

The majority of content words not listed in the dictionary are tagged using morphological information about the suffixes, in the morpho-conceptual lexicon, as illustrated with the following partial entries.

(26) a. -al: Th (x), PRED (e), RESULT (e)
Prop (x), MOD (r), PROPERTY (r)

b. -er: Th (x), PRED (e), AGENT/INST(e)
$Prop_F$ (x,y), Op(y), MORE PROP(x)than (y)

Disambiguation is performed by analyzing local morpho-syntactic context using a bottom-up chart parser and a Unification grammar. The latter form constituents and compositionally derive the meaning of expressions, from the meaning of their parts. In our IP model, every module of the linguistic processor is based on formal and semantic asymmetries.

We expect further research and experiments to show that the performance of IP systems incorporating (23) increases, with respect to both the accuracy of the information extracted from texts as well as with respect to the relevance of the documents retrieved.

4. Conclusion

We presented the main features of an integrated theory of language knowledge and use and provided evidence from morphological as well as syntactic structure for the plausibility of such a model. Our integrated model keeps maximally simple the relation between the grammar and the parser while it keeps constant the specificity of each system. The integration of asymmetry-based parser in IP systems is expected to improve their performance.

References

Anderson, S. (1992) A-Morphous Morphology. Cambridge: Cambridge University Press.

Antworth, E. (1990) PC-KIMMO: A Two Level Processor for Morphological Analysis. Dallas, TX: Summer Institute of Linguistics.

Arampatzis, A.T., T. Tsoris and C.H.A. Koster, 1997. IRENA: Information Retrieval Engine Based on Natural Language Analysis. RIAO 97 Proceedings, McGill University.

Berwick, R. 1991. Principles of Principle-Based Parsing. In R. Berwick, S. Abney & C. Tenny (eds.) Principle-Based Parsing. Dordrecht: Kluwer.

Boguraev, B. & J. Pustejovsky (eds.) 1996. Corpus Processing for Lexical Acquisition. Cambridge, Mass.: MIT Press.

Chomsky, N. 1981. Lectures on Government and Binding. Dordrecht: Foris.

Chomsky, N. 1986. Barriers. Cambridge, Mass.: The MIT Press.

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

Chomsky, N. 1998. Minimalist Inquiries. Ms. MIT.

Church, K.1988. Stochastic Parts Program and NP Parser for Unrestricted Text. Proceedings of the Second Association of Computational Linguistics Conference on Applied Natural Language Processing.

Cinque, G. 1997. Adverbs and Functional Heads. A Cross-Linguistic Perspective. Oxford University Press.

Copestake, A., & T. Briscoe. 1996. Semi-productive Polysemy and Sense Extension. In J. Pustejovsky & B. Boguraev (eds.) Lexical Semantics. The Problem of Polysemy. Oxford University Press.

Derose, S. J. 1988. Grammatical Category Disambiguation by Statistical Optimization. Computational Linguistics, 14.1.

Di Sciullo, A.M. 1996. Modularity and Xo/XP Asymmetries. Linguistic Analysis 26.

Di Sciullo, A.M.. 1997a. Prefixed verbs and Adjunct Identification. In A.M. Di Sciullo (ed.) Projections and Interface Conditions. Oxford University Press.

Di Sciullo, A.M. 1997b. Argument Structure Parsing. In A. Ralli, M. Grigoriadou, G. Philokyprou, D. Christodoulakis and E. Galiotou eds. Papers in Natural Language Processing.

Di Sciullo, A.M. 1998. Features and Asymmetrical Relations in Morphological Objects. GLOW 1998 Newsletter.

Di Sciullo, A.M. 1999. The Local Asymmetry Connection. MIT Working Papers in Linguistics. Cambridge, MASS. MIT.

Di Sciullo, A.M. & C. Tenny. 1997. Modification, Event Structure and the Word/Phrase Asymmetry. NELS 18.

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

Di Sciullo, A.M. and C. Thérien, in progress Asymmetry-Based Stemming.

Door, B. 1991. Principle-Based Parsing for Machine Translation. In R. Berwick, S. Abney & C. Tenny (eds.) Principle-Based Parsing. Dordrecht: Kluwer.

Epstein, S. 1995. The Derivation of Syntactic Relations. Ms. Harvard University

Fong, S. The Computational Implementation of Principled-Based Parsers. In R. Berwick, S. Abney & C. Tenny (eds.) Principle-Based Parsing. Dordrecht: Kluwer.

Halle, M. and A. Marantz, 1993. Distributed Morphology and the Pieces of Inflection. In K. Hale and J. Keyser (eds), The View from Building 20. Cambridge, MA: MIT Press

Kartunnen, L. 1983. KIMMO: A General Morphological Processor. Texas Linguistic Forum 22.

Kashket, M. 1991. Parsing Walpiri. A Free Word Order Language. In R. Berwick, S. Abney & C. Tenny (eds.) Principle-Based Parsing. Dordrecht: Kluwer.

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

Marken, C.G. 1990. Parsing the LOB Corpus. Association of Computational Linguistics Annual Meeting.

Pohlmann R, and W. Kraaij, 1997. The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts. RIAO 97 Proceedings. McGill University.

Porter, M.F. 1980. An Algorithm for suffix Stripping . Program, 14.3.

Pustejovsky, J., B. Boguraev, M. Verhagen, P.Buitelaar, M. Johnston, 1997. Semantic Indexing and Typed Hyperlinking. In Natural Language Processing for the World Wide Web. Papers from the 1977 AAAI Spring Symposium. AAAI Press.

Robert, F. & K. Vijayashankar, 1995. C-command and Grammatical Primitives. GLOW 1995 Newsletter.

Reuland, E. 1998. Deriving C-command in Binding. NELS 18.

Savoy, J. 1993. Stemming of French Words Based on Grammatical Categories. Journal of the American Society for Information Sciences, 44.1.

Uszkoreit, H. 1990. Unification in Linguistics. Class Lectures, 2nd European Summer School in Language, Logic and Information. Leuven.