BRIDGING THE GAP BETWEEN LINGUISTIC THEORY AND NATURAL LANGUAGE PROCESSING

Bento Carlos DIAS-DA-SILVA

UNESP/FAPESP
Rodovia Araraquara-Jaú, Km 1
Araraquara, SP - CEP 14800-901 - Brasil

Abstract: This paper addresses the problem of how to bridge the gap between linguistic theory and natural language processing research. It characterizes an approach to natural language processing based on cooperative work between different specialists, in particular, between linguists and system designers, each team working out problems in a specific phase of development in tandem. We suggest that natural language processing systems are a particular type of knowledge systems where a cluster of linguistic and extra-linguistic knowledge is represented and applied electronically to exploit natural language tasks such as spelling checking, building of grammars and lexicons, machine translation, and natural language understanding and generation.

Keywords: natural language processing research, linguistic theory, knowledge processing systems, knowledge engineering.

1. INTRODUCTION

It is undeniable that the ultimate challenge for natural language processing system designers has been to develop computational systems that are capable of handling man-machine communication by means of natural languages. But the achievements have been quite more modest. Most natural language processing research has been focusing on the computational processing of the orthographic forms of natural language utterances and text. The issues in speech production and recognition have been much more difficult to settle down.

Furthermore, there have been drawbacks, some of which due to either lack of appreciation for the complexity of natural languages or underspecification of the complexity of the task itself, which reveals a disturbing gap between NLP research and linguistic theory.

ISBN: 0 08 043 438X

To reduce the disturbing gap between "language scientists" and "language engineers", this paper will propose a unified framework to natural language processing studies to graduate students and researchers in Linguistics and related disciplines. We will show that natural language processing systems can be conceived of as a particular type of knowledge processing system where the complex of linguistic and extra-linguistic knowledge is elicited, represented, and applied electronically to exploit and to perform natural language tasks. Accordingly, this approach to natural language processing research program claims that the grammatical and discourse phases of processing should be tackled in three broad domains: (i) linguistic elicitation, (ii) computer representations and algorithms, and (iii) system building.

In Section 2, we overview the problems that contribute to widen the gap between "language scientists" and "language engineers", in Section 3, we present the overall three-domain approach to natural language processing research, and, in Section 4 we illustrate the dynamic of the proposal by applying it to the computational modelling of a very simple syntactic rule.

2. THE GAP

It is a fact that an overwhelming growth in the field of natural language processing (henceforth NLP) has taken place since the potential for building computer models of natural language text understanding and generation was recognized by the pioneers of machine translation, who struggled to try to tame natural languages in the early 1950's (Carbonell and Hayes, 1990; Gardner *et al.*, 1981; Gevarter, 1984). But it is also a fact that a multiplicity of projects has sprawled since then. As a result, NLP seems to be a discipline in ferment, which gathers researchers with a wide range of backgrounds and interests, emphasizing its diverse aspects, and employing manifold methods and techniques to build a number of varied commercial applications.

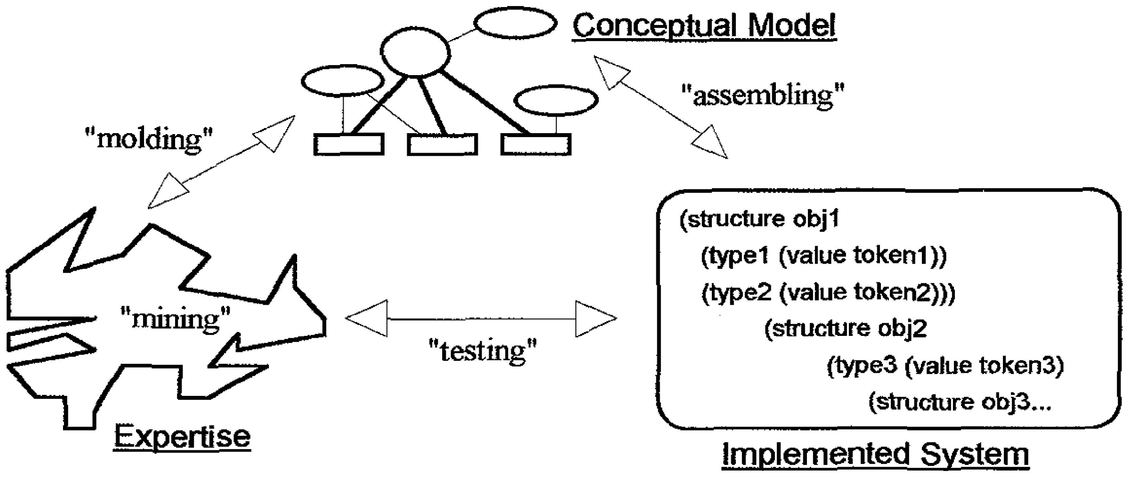
It is undeniable that the ultimate challenge for NLP system designers has been to develop computer programs that are capable of handling man-machine communication by means of natural languages. Despite the enthusiasm, the achievements, however, have been quite more modest. Most NLP research has been focusing on the development of very particular computational programs that perform very specific linguistic tasks such as spelling checking, hyphenation, dictionary look up, parsing, production and generation of "canned texts". The issues in speech production and recognition have even been barely touched.

On the one hand, it is not difficult to spot NLP projects that either resort to traditidional grammar only or strive to succeed without any recourse to linguistic theory (Winograd, 1972; Moreno Fernández, 1990); on the other hand, linguistic theory has either disregarded computational issues altogether or provided the ammunition to deaden the enthusiastic development of NLP technologies (Halvorsen, 1989; Mykowiecka, 1991). In addition, those who are new to either field have to confront an astounding number of technical reports, journal articles and conference papers, just to get acquainted with a number of approaches, or to decode a number of puzzling formalisms.

It is thus a fact that team work is called for (Sanders and Sanders, 1989; Bailin and Levin (1989); Searle, 1990; Starosta, 1991). In order to foster cooperative work particularly between linguists and NLP system designers, it is urgent to characterize an integrated approach to both NLP research and NLP system building.

3. THE THREE-DOMAIN APPROACH

Our starting point is the conception of knowledge systems - computer programs that store and apply specific knowledge to solve specific problems. Hayes-Roth (1990) argues, for instance, that knowledge, like a rare metal, lies dormant and impure, beneath the surface of consciousness. Once extracted, he says, an element of knowledge must undergo several transformations before it can be modelled. He finally concludes that the process of building knowledge systems requires that the knowledge system team performs four basic types of functions: "mining, molding, assembling, and refining knowledge." Figure 1 illustrates this style of development *(Durkin, 1994).*



```
(structure obj1
  (type1 (value token1))
  (type2 (value token2)))
        (structure obj2
              (type3 (value token3)
                  (structure obj3...
```

Fig. 1 Expert system development.

Accordingly, Dias-da-Silva (1996) demonstrates that NLP systems can be conceived of as a particular type of knowledge processing systems where the complex of linguistic and extra-linguistic knowledge is represented and applied electronically to exploit and to perform a number of linguistic as well as metalinguistic tasks such as "check" spelling and grammar, "analyze" morphological and syntactic structures, "understand" and "produce" texts, "translate" words, sentences and texts, "make" and "answer" questions, and "help" linguists develop their own linguistic models, just to mention the most impressive ones (Nirenburg *et al.*, 1992).

ISBN: 0 08 043 438X

Thus, both NLP research programs and the task of building a particular NLP system should mirror and benefit from the strategies developed in the field of "knowledge engineering." Table 1 shows that if we are to model bits of both the linguistic competence and performance that speakers have of an individual natural language, we must specify their linguistic knowledge and abilities, and encode the resulting specifications into computer programs (Schank and Riesbeck, 1981). In other words, the strategy includes the tasks of modelling the specific types of knowledge that a natural language expert has, how this knowledge is acquired, stored, and applied.

Table 1 Tasks and basic resources in NLP development

Tasks		Basic Resources
	"MINING"	
● What knowledge and use of language should be worked out?		● Linguistic Theories of Competence and Performance
	"MOLDING"	
● How should the linguistic information be formally represented?		● Formal Representation Languages
	"ASSEMBLING"	
● How should the representations be encoded ?		● Programming Languages and Computer Systems

This amounts to saying that the process of designing and implementing a computer program that is capable of processing natural language should comprise at least three iterative and evolutionary phases of analysis in three complementary domains: (i) **Linguistic Domain** – With the recourse to explicit models proposed within the linguistic theory (Bresnan, 1982; Halliday, 1985; Gazdar, *et al.*, 1985; Chomsky, 1986; Barton, Berwick and Ristad, 1987; Sells, 1985; Kayser, 1989; Jackendoff, 1990; Chierchia and McConnell-Ginet, 1990; Cann, 1993), the task here is to gather a body of knowledge about the linguistic phenomena one is trying to model; (ii) **Representational Domain** – In this domain, the system conceptual design is pursued. It involves the selection and/or the proposal of computational tractable representation systems whose degree of expressiveness is powerful enough to encode the body of knowledge constructed in the previous phase (Minsky, 1975; Brachman and Levesque, 1985; Grosz, Jones, and Webber, 1986; Allen, 1987; Reyle and Rohrer, 1987; Pustejovsky and Boguraev, 1991; Partee *et al.*, 1993); and (iii) **Implementational Domain** – Finally, the previous formal encoding is "translated" into a suitable programming language and the overall system planning is developed (Clocksin and Mellish, 1987; Pereira and Shieber, 1987; Gazdar and Mellish, 1989; McCord, 1990). That is, the conceptual components of the system are assembled and the computational environment in which the natural language system is to be developed and implemented is designed.

It should be noted that the fourth function aforementioned, "refining", is not considered here as a fouth domain because it refers to the testing of the constructs proposed within each of the three domains. It is more appropriately considered as a checking phase where the products of the whole cycle – elicitation, formalization, and implementation –, or parts of it, are revised.

Figure 2 shows that the three domains can in turn be re-interpreted as a three-phase cycle in the process building a particular NLP system: (i) elicitation of linguistic knowlwdge and usage ("mining"), (ii) representation of that knowledge ("molding"), and (iii) encoding of the resulting representations into computer programs ("assembling").

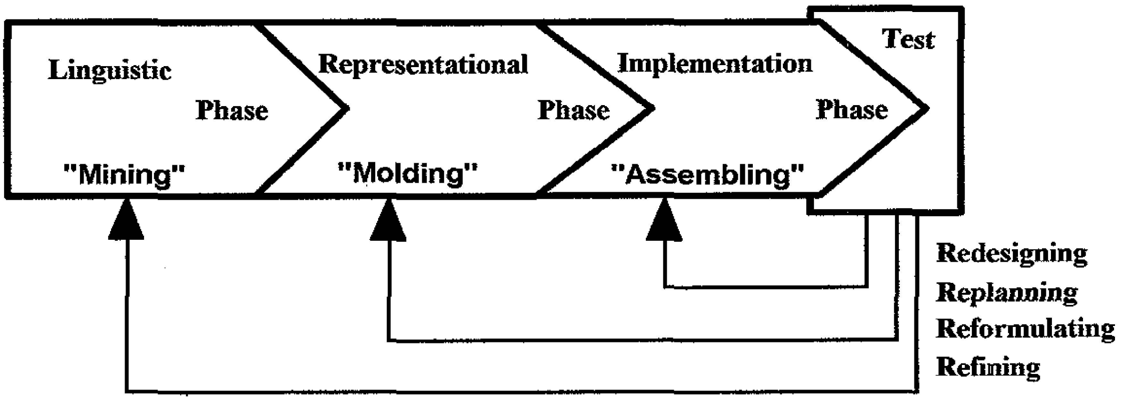


Fig. 2 The three-domain approach to NLP.

In other words, NLP system building requires the computational encoding of a considerable number of constructs that encodes the linguistic competence-performance information the system is intended to model: the linguistic knowledge and usage elicited during the linguistic phase is transformed into computationally tractable constructs in the representation phase, and such constructs are further encoded in computational programs. Table 2 synthesizes the natural language processing research cycle and its target "products".

Table 2 The NLP research cycle.

Tasks	Products
• elicitation of linguistic knowledge	• linguistic representations
• formalization of linguistic knowledge	• computational tractable representations
• implementation of constructs	• computer programs

ISBN: 0 08 043 438X

4. AN ILLUSTRATION

Figure 3 shows the three-phase process of representing bits of domain-specific knowledge necessary to modelling an English basic sentence structure.

LINGUISTIC PHASE	**LINGUISTIC DESCRIPTION**
	An English simple sentence structure is described by the concatenation of a noun phrase and a verb phrase. The noun phrase is the subject and the verb phrase is the predicate. The subject and the verb must have the same number and person features. The grammatical case of the subject is nominative and the verb is finite. (Quirk & Greenbaum, 1977)
	LFG **FORMALISM**
	S → NP VP (↑SUBJECT)=↓ ↑=↓ (Bresnan, 1982)
REPRESENTATIONAL PHASE	***PATR*** **REPRESENTATION**
	Syntactic Rule: S → NP VP Features: < NP person > = < VP person > < NP number > = < VP number > < NP case > = nominative < VP verbal form > = finite (Shieber, 1986)
IMPLEMENTATION PHASE	***PROLOG*** **IMPLEMENTATION**

	`s(PO,P):-`
	`np(Person,Number,Case,PO,P1),`
	`vp(Person,Number,Case,P1,P).`
	(Clocksin & Mellish, 1981)

Fig. 3 A phrase structure rule.

5. CONCLUSION

One particular point should be mentioned concerning the people involved in an NLP system project. The main players on an NLP system project are undoubtely the language scientists and the language engineers. Each plays a key role in the development of the system. Accordingly, important qualifications are needed by each team to contribute effectively to the success of an NLP projetct. Language scientists are the ones that have expert knowledge of language, can communicate the knowledge, can aid in knowledge acquisition, and can help define interface specifiactions. Language engineers are the ones that have knowledge engineering and programming skills, can match problem to software, and can aid in system development. Both must have efficient problem-solving skills.

Finally, it should be pointed out that molding a team into a form where it can be productive in the challenging task of building NLP systems takes time and effort. Our goal has been to foster cooperative work between system designers and linguists. In particular, we have been concentrating our efforts on issues concerning phases (i) and (ii). Despite reflecting biases of research efforts, we have tried to present a unified framework to students and researchers in linguistics and related disciplines whose concerns include tackling the fascinating computer approach to the understanding of natural languages.

REFERENCES

Allen, J.F. (1987). *Natural language understanding*. Benjamin Cummings, Menlo Park.

Bailin, A. and L.S. Levin (1989). Introduction: Intelligent Computer-Assisted Language Instruction. ***Computers and The Humanities*** **23**, 1-2.

Barton, G.E., R.C. Berwick and E.S. Ristad (1987). *Computational complexity and natural language*. MIT Press, Cambridge, Mass.

Brachman, R.J. and H.J. Levesque (1985). *Readings in knowledge representation*. Morgan Kaufmann, San Mateo.

Bresnan, J. (Ed.) (1982). ***The mental representation of grammatical relations***. MIT Press, Cambridge, Mass.

Cann, R. (1993). *Formal semantics*. Cambridge University Press, Cambridge.

Carbonell, J.G. and P.J. Hayes (1990). Natural-Language Understanding. In: *Encyclopedia of artificial intelligence*, (E. Shapiro (Ed.)), pp. 660-77. Wiley, New York.

Chierchia, G. and S. McConnell-Ginet (1990). ***Meaning and grammar***. MIT Press, Cambridge, Mass.

Chomsky, N. (1986). *Knowledge of language: its nature, origins, and use*. Praeger, New York.

Clocksin, W.F. and C.S. Mellish (1987). ***Programming in prolog***. Springer-Verlag, Berlin.

Dias-da-Silva, B.C. (1996). The technological facet of language studies: natural language processing, PhD diss, FCL-UNESP, Araraquara, Brasil.

Durkin, J. (1994). *Expert systems: design and development*. Prentice Hall International, London.

ISBN: 0 08 043 438X

Gardner, A. *et al.* (Eds.) (1981). Understanding Natural Language. In: *The handbook of artificial intelligence* (A. Barr and E.A. Feigenbaum (Eds.)), pp. 224-321. William Kaufmannn, Los Altos.

Gazdar, G. and C. Mellish (1989). ***Natural language processing in prolog: an introduction to computational linguistics***. Addison-Wesley, New York.

Gazdar, G. *et al.* (1985). *Generalized phrase structure grammar*. Basil Blackwell, Oxford.

Gevarter, W.B. (1984). *Artificial intelligence, expert systems, computer vision, and natural language processing*. Noyes, Park Ridge.

Grosz, B., K. Jones and B. Webber (Eds.) (1986). *Readings in natural language processing*. Morgan Kaufmann, Los Altos.

Halliday, M.A.K. (1985). ***An introduction to functional grammar***. Edward Arnold, London.

Halvorsen, P-K. (1989). Computer Applications of Linguistic Theory. In: *Linguistics: the cambridge survey II* (F. Newmeyer (Ed.)), pp. 198-219. Cambridge University Press, Cambridge.

Hayes-Roth, F. (1990). Expert Systems. In: *Encyclopedia of artificial intelligence* (E. Shapiro (Ed.)), pp. 287-98. Wiley, New York.

Jackendoff, R (1990). *Semantic structures*. MIT Press, Cambridge, Mass.

Kayser, H. (1989). Some Aspects of Language Understanding, Language Production, and Intercomprehension in Verbal Interaction. In: *Connexity and coherence* (W. Heydrich *et al.* (Eds.)), pp. 342-65. Walter de Gruyter, Berlin.

McCord, M. (1990). Natural Language Processing in Prolog. In: *Knowledge systems in prolog* (A. Walker *et al.*), pp. 337-450. Addison-Wesley, Reading.

Minsky, M. (1975). A Framework for Representing Knowledge. In: *Mind design* (J. Haugeland (Ed.)), pp. 95-128. MIT Press, Cambridge, Mass.

Moreno Fernández, F. (1990). Lingüística Informática e Informática Lingüística. *Lingüística Española Actual* 12, 5-16.

Mykowiecka, A. (1991). Natural-language generation – an overview. *International Journal of Man-Machine Studies* 34, 497-511.

Nirenburg, S. *et al.* (1992). *Machine translation*. Morgan Kaufmann, San Mateo.

Partee, B.H. *et al.* (1993). *Mathematical methods in linguistics*. Kluwer, Dordrecht.

Pereira, F.C.N. and S. Shieber (1987). *Prolog and natural language analysis*. Chicago University Press, Chicago.

Pustejovsky, J. and B. Boguraev (1991). Lexical Knowledge Representation and Natural Language Processing. *IBM Journal of Research and Development* 35, 1-20.

Quirk, R. and S. Greenbaum (1977). *A university grammar of English*. Longman, London.

Reyle, U. and C. Rohrer (1987). *Natural language parsing and linguistic theories*. D.Reidel, Dordrecht.

Sanders, A. and R. Sanders (1989). Syntactic Parsing: A Survey. ***Computers and The Humanities* 23**, 13-30.

Schank, R.C. and C.K. Riesbeck (Eds.) (1981). *Inside computer understanding*. Lawrence Erlbaum, Hillsdale

Searle, J.R. (1990). Foreword. In: *Reference and computation* (A. Kronfeld), pp. xiii-xvii. Cambridge University Press, Cambridge.

ISBN: 0 08 043 438X

Sells, P. (1985). *Lectures on contemporary syntactic theories:* €hicago University Press, Chicago.

Shieber, S.M. (1986). *An introduction to unification-based approaches to grammar*. Chicago University Press, Chicago.

Starosta, S. (1991). Natural Language Parsing and Linguistic Theories: Can The Marriage Be Saved?. *Studies in Language* 15, 175-97.

Winograd, T. (1972). *Understanding natural language*. Academic Press, New York.