Preface

This text is intended to give a general view of the IBM-INESC Scientific Group, presenting its history, organization, human resources, technical results, list of documentation produced, and some external opinions on its achievements.

Its main goal is to prove that not only the environment but also the technical and organizational decisions taken are adequate to develop Natural Language Processing, and specifically machine translation, in Portugal.
Table of Contents

General description .................................................. 1
Short history ........................................................... 3
Work in progress ....................................................... 5

Organization ............................................................ 7
  Distribution of the work among the several members ............... 7
  Coordination of work ............................................... 7
    Meetings ................................................................ 7
    Flow of information chain ........................................ 7

Human resources ......................................................... 9
  Organizers ................................................................ 9
  Base workers ................................................................ 9
  Previous members ....................................................... 10
    Short curricula ....................................................... 11
      Ana Maria Ferreira Pereira dos Santos ....................... 11
      António Manuel Maduro Colaco ................................ 11
      Carlos Miguel Marques Correia ................................ 11
      Dalila Maria Garcia Rosales ...................................... 12
      Diana Maria de Sousa Marques Pinto dos Santos ........... 12
      Jorge Manuel de Almeida Sietra ................................ 12
      José Campos de Azevedo ......................................... 13
      José Carlos Dinis Medeiros ...................................... 13
      José Fernando Cabeleira ......................................... 13
      José Nuno Silva Costa Alves dos Santos ...................... 14
      José Virgilio Moreira Fragoso ................................... 14
      Júlio António dos Reis Rodrigues ............................... 14
      Maria Eduarda de Sousa Costa ................................... 15
      Maria Filomena Martins Castanheira Amorim Galvão ....... 15
      Maria Regina Caldeira dos Reis ................................ 15
      Paulo Jorge Viegas Fernandes ................................... 15
      Rui Pedro Ribeiro Marques ....................................... 16
      Teresa Maria Brás de Brito Pires ............................... 16

Use and development of the machine translation system ............... 17
  Invoking the translator .............................................. 17
  Dictionaries ............................................................ 17
  Programs ................................................................. 18
  Debugging ................................................................ 18
  General utilities ...................................................... 18
  System organization .................................................. 19
  Machines .................................................................. 19
  Files ...................................................................... 19

Documentation ................................................................ 21

Present status and near future ............................................. 23
  Linguistic overview ................................................... 23
  Bilingual dictionary ................................................... 23
  System design .......................................................... 24
  Actual results ........................................................... 24
General description

This Group was created as a research environment linking two Portuguese institutions, IBM Portugal and INESC (Instituto de Engenharia de Sistemas e Computadores).

Its first project was in the field of Natural Language Processing, namely in machine translation from English to Portuguese, integrated in the IBM project MENTOR. The work in this area continues to evolve, and significant results can already be displayed in the present day.

The writing of a grammar for Portuguese began later and was at once integrated in the overall system, providing thus a powerful programming environment for natural language processing as a whole.

The choice of research in the particular area of Natural Language Processing for the first tasks of the Group originated from the opinion that Portugal is the best place to do research and develop know-how in the automated handling of the Portuguese language, which has, by the way, received significantly less attention than other major languages.

At this moment, we can state the existence of a working group in NLP that can be considered the Natural Language Group in INESC, and which is respected as producing quality work by other NLP groups abroad both within IBM and in general.

---

1 MENTOR is an internal project name, standing for Multitarget ENGLISH Translator, joining Haifa Scientific Center, Israel, Madrid Scientific Center, Spain, the Research Unit for Computational Linguistics of the University of Helsinki in a joint project with IBM Finland, and our Group.
Short history

In the 1st July 1987 a protocol between INESC and IBM Portugal was signed, that established collaboration of the two institutions for joint research, and the MENTOR translation project was mentioned as first project. Two people were hired by INESC for this task.

The managers responsible for this process were Jorge Ferreira Pinto from IBM and Luis Vidigal from INESC.

These were the most significant dates of the short history of the Group:

**July 1987**  
Stay in Madrid Scientific Center for a first contact with the tools that were to be used by MENTOR, and with machine translation in general.

**November 1987**  
PLNLP/PEG course in Hawthorne, IBM Yorktown Heights, to learn the programming language, and how to use and make the best of the English parser PEG.

**February 1988**  
First proposal on MT architecture, broadcast to the other groups

**July 1988**  
First and only MENTOR meeting, in Lisbon.

The Portuguese prototype was presented to the other teams and to Victor Martinez, of IBM Headquarters in Paris.

**October 1988**  
Presentation of a paper on MENTOR88/P in the Thornwood conference IBM-ITL.

**November 1988**  
PLNLP grammar course in Bethesda; beginning of the Portuguese grammar.

**February 1989**  
Significant increase of the size of the group: the number of full time members rose from three to six.

**March 1989**  
Paper submitted to EP1A-89 (AI conference in Lisbon).

Suggestion by Stephen Richardson, ASD Bethesda, that one idea in this paper should be patented.
Work in progress

In the general field of Natural Language Processing, there is at this moment work being done in the following domains:

- Machine translation (linguistics)
  - Contrastive syntax between English and Portuguese
  - Tense and aspect representation
- Machine translation (organization)
  - Integration with the grammar for multiword expressions
  - Style processing, or how to produce multiple equally valid translations.
  - Generation as a separate module
- Dictionaries
  - Building of an English to Portuguese MT dictionary
    - Terminology
    - Lexical transfer
    - Syntactical markers fulfilling
- Grammar
  - Development
  - Definition of the syntactical features that should be in the dictionary (the actual filling will be done outside)
  - Documentation
- Connections with the exterior
  - Conferences support (for instance, Jaime Carbonell from CMU in July 1988 and Jonh Sowa next September).
  - Demonstration of the system to NLP experts visiting the Group (for instance Cristina Marques from USP (Brasil), Gabriel Lopes from UNL (Portugal), Miguel Filgueiras from UP (Portugal), Jens Aage Bertelson (IBM Denmark), Jaime Carbonell from CMU (USA))
  - Master's dissertation written within the group.
  - Internal presentations on NLP matters, rotatively organized by the members of the Group.
Organization

The increased number of people working in our system towards its improvement has brought forward several challenges as far as organization is concerned.

Not only the computational and physical space resources are scarce and have to be shared, but also the interaction of several people with different backgrounds (namely in Linguistics / Arts or Computer Science / Engineering) and producing parallel work has to be taken into account, in order to ensure an optimized merge of each member’s work.

Organization matters can be seen along three axes:

- distribution of the work among the several participants
- general working chain inside the Group
- technical support for the merge of parallel work

This last point will be described later in “System organization” on page 19.

Distribution of the work among the several members

There are two categories of workers in the Group: those with some organization responsibilities (organizers) and those with specific linguistic or programming tasks only (base workers).

The latter have one definite problem to handle, on which they are supposed to produce a report and/or the actions to solve it (be they in terms of translation specifications or actual programs).

The former have a more general area of work assigned, that they supervise and distribute among base workers, and work also in more englobating problems.

Which particular tasks are being worked on and who performs them is described in “Human resources” on page 9.

Coordination of work

Since there are both linguists and computer scientists working on this same project, it is necessary an interface between the two teams. And we try also to create a cooperative atmosphere among all members within the Group.

Meetings

Once a fortnight there is a meeting with each team so that they can present their doubts and talk about the development of their work. New improvements of the whole system or general problems found are also announced during the reunion.

If an extra meeting is needed, that information is broadcast to all machines.

Flow of information chain

So that a particular translation problem be solved, these steps are to be followed in our Group:

---

2 It should be emphasized that several processes of this kind take place in parallel.
1. A linguist studies a specific problem and elaborates a report.
2. The report is studied and discussed with Regina Reis.
3. Ana Santos analyses it to see if the rules presented can be programmed.
4. If everything is right, one can go to next step; otherwise the report must be reformulated.
5. As soon as possible, programming collaborators program the rules. They must create a file with some sentences concerning the problem in question, in order to verify if they are well translated.
6. When they consider their task done, it is checked by Nuno Alves. One can go to next step unless the programmed rules are not right. In this case they must be written again.
7. When the rules are ready, they are sent to the linguist who specified them. It is his/her turn to check them in order to testify that the problem he/she wanted to solve is indeed solved. If not, a respecifying of the linguistic rules is needed, and the process begins from point 1 above. Otherwise, the linguist proceeds to write the second half of his/her report, where the tests performed and the eventual problems not yet solved (if any) should be described. The process is thus ended for that particular problem, in what the linguist is concerned.
8. The rules are then again sent to Nuno Alves. When there is already a significant number of new rules, a new translator should be built. It is Nuno's task to check whether the independant new rules, when joined together, do not interact in harmful ways, degrading the system's performance. If this happens, the computer scientists who wrote those rules must gather to solve the problem.
9. After the new environment has been created and extensively tested, it takes the place of the old one.
Human resources

The Group members are very diverse, not only in background but also in kind of contract (and consequently, of duration of work) that they have with the Group.

At present, IBM Portugal pays INESC 5 graduate scholarships plus one undergraduate scholarship, which corresponds to 2 full-time workers, 2 part-time workers, 5 small-time workers (one-fourth of full time), 1 full-time student, and 7 students on a free-time schedule. One IBM employee is moreover assigned at INESC.

The distribution of the work and responsibility among the members is sketched briefly in the following paragraphs. A short curriculum of each member ends this presentation.

Organizers

Diana Santos

Main supervision.

In charge of the Portuguese grammar (for confidentiality reasons, it was not yet possible to give explicit responsibilities to other members).

Working on translation by multiword expressions, recurring to the Portuguese grammar.

Regina Reis

Supervision of the linguist collaborators.

Main area of research: tense and aspect transfer.

In charge of dictionary supervision.

Ana Santos

Control of the work that is given to the computer scientists, both by checking the specifications produced by the linguists and by distributing them.

In charge of the transfer phase.

Main research task: accommodate several stylistic possibilities for translation (at the moment, only the first is displayed/used).

Nuno Alves

Supervision of the computer scientist collaborators, in their actual tasks.

In charge of the generation phase.

Main research task: significant improvement of generation capabilities.

Base workers

José Carlos Medeiros

Development of tools (in REXX) for user friendly environment, concerning both dictionaries and test and documentation facilities.

Study of evaluation matters.
António Colaço

Reflexivization in English and its correspondence in Portuguese. Study of Portuguese reflex constructions. Transfer rules to handle these last, resulting in a different target structure.

Study of English complex noun phrases and their transfer to Portuguese.

José Azevedo

Classification of conjunctions according to the verbal mode they require, and their result in the target sentence. Study of adjectives and adverbs that may have similar consequences, for instance implying a change from indicative to subjunctive or infinitive mood.

Virgílio Fragoso


Study of Portuguese defective verbs and how to handle them in translation.

Filomena Galvão

Study of the occurrence of adjectives with the verbs “ser”, “estar”, and both.

General subcategorization of adjectives - an overview.

Dalila Rosales

Prepositional verbs: study of their correspondence in Portuguese and the syntactical transformations they imply. Storing, in the bilingual dictionary, the corresponding preposition or argument structure necessary for the translation.

Eduarda Costa

Test and study of the system from a translator’s point of view.

Statistical studies of use of verbs and tenses.

Comparison with human-translated text. This member is a translator specialized in English Computer Science texts.

Rui Marques

Study of quantifiers; differences of their behavior in Portuguese and English; their agreement requirements in Portuguese. Start with numerals.

Teresa Pires

Tense transfer: study of past perfect. Elaboration of rules to translate this tense into Portuguese, in simple and compound sentences, including the study of its correspondence in subjunctive mood.

João Cabeleira

Jorge Sietra

Paulo Fernandes

Júlio Rodrigues

Carlos Correia

Programming of the linguistic rules in PLNLP.

Previous members

Though the Group is recent, there have been already some people who left it after successfully accomplishing work for the machine translation project.

Paulo Libano Monteiro

10 IBM-INESC Scientific Group
One of the two initial members of the Group. He worked here during the first year, having left in August 1988.

He was the main author of the generation phase and of tense transfer in the MENTOR88/P prototype.

Fátima Pedroso
The first full-time undergraduate student in the Group, in a 5-months assignment (October 1988 to March 1989), she was then replaced by José Carlos Medeiros.

She is the only author of the program TRADUZIR (see "General utilities" on page 18) whose technical documentation she is still finishing.

Short curricula

Ana Maria Ferreira Pereira dos Santos

Date of birth: 1963/02/18

Type of collaboration within the Group: She started working on February 1989, as a full time collaborator.

Academic and professional training:
• Graduated by I.S.T.(Instituto Superior Técnico), Universidade Técnica de Lisboa, on Electrotechnic Engineering (in the area of Systems and Computers). Her degree was completed on December, 1988.
• The final project of the university course was developed at INESC. It concerned the design of a centralized computer network and it was implemented in Turbo Pascal.
• Knowledge of English and French.

António Manuel Maduro Colaço

Date of birth: 1961/01/12

Type of collaboration within the Group: He began working on March 1989, as a half time collaborator.

Academic and professional training:
• High School Diploma by the Danbury High School, Connecticut, USA.
• Cambridge Certificate of Proficiency in English by the British Institute in Portugal.
• Graduated by the Faculty of Arts of the Lisbon University in Portuguese and English Linguistics and Literatures.

Main occupation: Teacher of Portuguese Language and Literature in secondary school.

Carlos Miguel Marques Correia

Date of birth: 1967/09/12

Type of collaboration within the Group: He began working in March 1989, as a free-time student (estimated 6 hours / week).

Academic and professional training:
• At present attending the fourth year of Electrotechnical Engineering, option of Systems and Computers, in I.S.T. (Instituto Superior Técnico), Universidade Técnica de Lisboa.
• Working knowledge of the following operating systems: VM/SP, VAX/VMS, UNIX and MS/DOS.
• Working knowledge of the following programming languages: Pascal, Basic, C, Fortran and PL/NP.

**Dalila Maria Garcia Rosales**

*Date of birth:* 1965/03/29

*Type of collaboration within the Group:* She started working in March 1989, 10 hours / week.

*Academic and professional training:*

• She graduated in "Línguas e Literaturas Modernas" (English / French) in 1987, at the Faculdade de Letras de Lisboa, Universidade Clássica de Lisboa.

*Main Occupation:* Teacher of English in a secondary school.

**Diana Maria de Sousa Marques Pinto dos Santos**

*Date of birth:* 1962/03/01

*Type of collaboration within the Group:* She began working in July 1987, as a part-time worker. In September 1988 she became a full time member.

*Academic and professional training:*

• She graduated in Electrotechnical Engineering (Telecommunications and Computers) at I.S.T. (Instituto Superior Técnico), Universidade Técnica de Lisboa, in 1985.

• She taught Electronics and Introduction to Circuits and Signals during her last two years as an undergraduate student, and was a teaching assistant in Applied Mathematics from 1985 to 1986 and in Introduction to Computer Science from 1986 to 1987.

• She was assigned as collaborator of the CIIST (I.S.T.'s Computing Center) from 1987 to 1988, till she left her place at the University to join IBM. During that period, she was in charge of providing education and support to languages used in Artificial Intelligence, such as LISP (Franz LISP) and OPS5. She wrote and taught a course on each language.

• She took her Master's degree in Electrotechnical Engineering (Telecommunications and Computers), at the same University, from 1985 to 1988. The main emphasis was in Artificial Intelligence. Specifically in Natural Language Processing she produced an OPS5 program for automatic accentuation of Portuguese text, a small ATN grammar for interacting in Portuguese with a semantic network (SNePS), and wrote her dissertation on one subset of the work for machine translation at the IBM-INESC Scientific Group.

• She worked at INESC from 1984 to 1987 in the Speech Processing Group, oriented by Prof. Borges de Almeida. In this environment she completed her final course project (an LPC vocoder on a TMS32010 microprocessor). She also worked on vector quantization of speech signals, and produced a Pascal program for performing this task.

• She was in charge of the AI module in an EEC course for undergraduated students, in 1987, that she organized and taught.

**Jorge Manuel de Almeida Sietra**

*Date of birth:* 1969/03/17

*Type of collaboration within the Group:* He began working on March 1989, as a free time student (estimated 6 hours a week).

*Academic and professional training:*

• He is a second-year student in Electrotechnical Engineering at the I.S.T. (Instituto Superior Técnico), Universidade Técnica de Lisboa.

• Three years graduation in Technical and Professional of Electronics, finished with an average of 16 (grades are 1 to 20).
• At present working as a part-time collaborator in the Computer Science Laboratory of the Mechanical Engineering Department of I.S.T..
• First Certificate of English from the University of Cambridge Local Syndicate.
• Working knowledge of the following programming languages: Pascal, Basic, C, PLNLP.
• Working knowledge of the following operating systems: YAX/VMS, MS/DOS, VM/SP.

José Campos de Azevedo

Date of birth: 1955/09/27

Type of collaboration within the Group: He began working on March 1989, 10 hours / week.

Academic and professional training:
• Graduated in Clinic Psychology by ISPA (Instituto Superior de Psicologia Aplicada), Universidade Técnica de Lisboa. Special interest in language acquisition and language disfunctions.
• First Certificate in English (British Institute).
• Complementary Course in Computer Science in the Ferreira Borges School.
  He developed two programs for Personality Testing in Basic, and has some experience in other programming languages, such as Pascal, COBOL and FORTRAN.

Main Occupation: English and Portuguese teacher in a preparatory school.

José Carlos Dinis Medeiros

Date of birth: 1965/08/13

Type of collaboration within the Group: He works since March 1989 as a full time collaborator, for a 5 months period, which will be an integral part of his graduation in Applied Mathematics.

Academic and professional training:
• He is doing his graduation in Applied Mathematics at U.B.I., Universidade da Beira Interior. He plans to graduate not later than this summer. As last-year project he worked in statistical quality control.
• He gave practical classes at his University in the subjects of Mathematical Analysis, Introduction to Computers and Basics of Graph Theory.

José Fernando Cabeleira

Date of birth: 1969/03/09

Type of collaboration within the Group: He began working on March 1989, as a free-time student (estimated 6 hours / week)

Academic and professional training:
• He is a second-year student in Mechanical Engineering at the I.S.T. (Instituto Superior Tecnico), Universidade Técnica de Lisboa.
• At present working as a part-time collaborator in the Computer Science Laboratory of the Mechanical Engineering Department of I.S.T..
• Three years course of Technical and Professional of Electronics, finished with an average of 15 (grades are 1 to 20).
• Working knowledge of the following programming languages: Pascal, C, Dbase, COBOL, BASIC, PLNLP.
• Working knowledge of the following operating systems: MS/DOS, VAX/VMS, CMS(VM/SP).

José Nuno Silva Costa Alves dos Santos

Date of birth: 1964/04/14

Type of collaboration within the Group: He began working on March 1989, as a part-time collaborator, (half of full time)

Academic and professional training:

• He is a last-year student of Electrotechnical Engineering and Computers at I.S.T (Instituto Superior Técnico), Universidade Técnica de Lisboa. He plans to get his degree next July.

• He has worked at INESC from 1986 to 1987 in the Electronic Commutating project, oriented by Prof. Guilherme Arroz and Eng. Pedro Faria Lopes. He worked on communication and compatibility between micro and mini-computers.

• From 1987 to the end of 1988; he was in INESC at the DICE project, oriented by Prof. José Delgado. DICE's goal is the design of a distributed parallel machine, where he has developed an intelligent hard disk interface, based on the Transputer microprocessor. He has, as well, developed some software, in OCCAM, for this interface.

• Still in DICE, he accomplished his final project (with a grade of 18 out of 20): he built an integrated graphic and text editor, for the DICE machine, written in an object oriented programming language, Smalltalk 80.

• Since September 1986, he has monitored educational actions financed by EEC. These actions covered programming in BASIC, Pascal, Lotus 1-2-3 and dBaseIII, and the use of operating systems, text processing, spreadsheets, graphic environments and databases.

José Virgílio Moreira Fragoso

Date of birth: 1958/3/25

Type of collaboration within the Group: He began working on March 1989, 10 hours / week

Academic and professional training:

• Degree in Education - Escola do Magistério Primário.

• Student of Linguistics at Faculdade de Letras de Lisboa. This course has also English and Italian as subjects of study.

Main occupation: Primary school teacher / pedagogical advisor: Supervisor of in-service training for junior colleagues

Responsibility for applying and integrating computer into the school curriculum, in a pilot experience.

Júlio António dos Reis Rodrigues

Date of birth: 1967/04/13

Type of collaboration within the group: He began working on March 1989, as a free-time student (estimated 6 hours / week).

Academic and professional training:

• At present attending the fourth year of Electrotechnical Engineering, option of Systems and Computers, in I.S.T (Instituto Superior Tecnico), Universidade Técnica de Lisboa.

• Working knowledge of the following operating systems: VM/SP, VAX/VMS, UNIX and MS/DOS.
• Working knowledge of the following programming languages: Pascal, Basic, C, PLNLP, Ada, Modula, Fortran.

**Maria Eduarda de Sousa Costa**

*Date of Birth:* 1932/02/11

*Type of collaboration within the Group:* She started working March 1989, 10 hours / week.

*Academic and professional training:*
  - Courses attended at Universidade Nova de Lisboa
    - Translation Science
    - Textual grammar of Portuguese
  - Four English Courses - Progressive levels - provided by a Public Administration Service
  - Good reading and writing in English and French.
  - Some understanding of German and Russian.

*Main Occupation:* Professional translator at the Public Administration, specialized in Computer Science Literature.

**Maria Filomena Martins Castanheira Amorim Galvão**

*Date of Birth:* 1961/01/16

*Type of collaboration within the Group:* She began working on March 1989, 10 hours / week.

*Academic and professional training:*
  - Student at the Faculdade de Letras de Lisboa, attending the course on Languages and Modern Literatures (Portuguese / English). She plans to get her degree by this summer.
  - First Certificate in English (Cambridge School).
  - Knowledge of French, German and Italian.

*Main Occupation:* Teacher of English and Portuguese at a preparatory school.

**Maria Regina Caldeira dos Reis**

*Date of birth:* 1950/09/27

*Type of collaboration within the Group:* She started working in September 1988, as a full time collaborator.

*Academic and professional training*
  - Graduated in "Línguas e Literaturas Modernas" (Portuguese / English) by the Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, in 1986.
  - First year of Master's in Linguistics - Syntax and Semantics, in Faculdade de Letras de Lisboa.
  - Participation in the first National Congress on Translation held by Instituto Superior de Línguas e Administração.
  - Knowledge of French and German.

**Paulo Jorge Viegas Fernandes**

*Date of birth:* 1967/09/27
Type of collaboration within the Group: He began working in March 1989, as a free time student (estimated 6 hours a week)

Academic and professional training:

- He is a second-year student in Electrotechnical Engineering at the I.S.T. (Instituto Superior Tecnico), Universidade Técnica de Lisboa.
- Finished the graduation in Electronics at School Emídio Navarro, with an average of 15 (grades are 1 to 20).
- Working knowledge of the following programming languages: Pascal, C, BASIC, COBOL, DBase, Fortran, PL/NL.
- Working knowledge of the following operating systems: VAX/VMS, MS/DOS, Unix, CMS(VM/SP).

Rui Pedro Ribeiro Marques

Date of Birth: 1969/12/22

Type of collaboration within the Group: He began working in March 1989, as a free-time student (estimated work load, 6 hours / week).

Academic and professional training: Student of linguistics (second year) at Faculdade de Letras de Lisboa, Universidade Clássica de Lisboa.

Teresa Maria Brás de Brito Pires

Date of birth: 1966/12/31

Type of collaboration within the Group: She began working in the Group on March 1989, as a free-time student.

Academic and professional training:

- She is a last-year student in “Línguas e Literaturas Modernas” (English / German) at the Faculdade de Letras de Lisboa. She plans to have her degree by this summer.
- She is attending the course on Translation (Portuguese / English) at the Centro Europeu de Línguas, which includes also training on MS/DOS, Wordstar and Lotus.
- She attended the sixth grade of the Goethe Institute for German.
- Knowledge of French.
Use and development of the machine translation system

A short description of how to use the machine translation (MT) system and which facilities are provided for its friendly development is attempted here.

This section is not supposed to furnish any detailed technical explanations, nor is it intended to give a comprehensive description of the tools and main programs that are in use in MENTOR. The reader is therefore invited to read the documents listed in "Documentation" on page 21 in order to know more about the whole system.

To complement the description of the management of human resources supplied in “Coordination of work” on page 7, some paragraphs will be devoted to the Group organization in what concerns technical matters. Special emphasis will be given to merge facilities and separability of work.

Invoking the translator

There are three different ways of invoking our MT system:

PORTUGA Interactively, inputting sentence by sentence and usually seeing the parse trees in English and in Portuguese, being able to access the dictionaries and inspect the internal structures, and to change the behavior of the translator (be it its display, or the syntactical sense that should be translated or even the actual translation of the English words).

PORT-RUN In batch mode, usually for a pertinent collection of sample sentences, to generally produce pairs English sentence / Portuguese sentence only (though the parse trees can be produced as well, if using ARV-RUN instead). This mode is generally employed for test purposes, for deep study of some particular problem, or for performance measures.

TRADUZIR In a totally independant way, just invoking a program on an English SCRIPT file and getting an equivalent Portuguese one. This mode is obviously not intended for system developers, but for end users, and at this moment is not really necessary yet.

Dictionaries

Associated with the translator, there are several dictionaries (OD dictionaries), all of them universally accessible. There is an English analysis dictionary which is supposed to give an almost unlimited coverage of American English; there is a bilingual dictionary under development; there is a Portuguese generation sparse dictionary3; and, finally, there is a Portuguese analysis dictionary being developed for the grammar, but that is used whenever the Portuguese parser is invoked for translation by MWEs.

To handle these dictionaries, which are easy to read and modify, we use the facilities provided by their authors, plus some others that we built on top of them, namely ODLISTA and ODFUNDE.

Any person can update its own version of the dictionary, stored in his/her machine. This comes in handy when different people are solving different problems in the same environment. To automatically merge several updates ODFUNDE (an extension of ODGEN) was designed.

3 as a provisional tool before the broad-coverage IBM internal lexicon is available for our purposes
Also, when making linguistic generalizations, it is necessary to take into account the classifications under which the English words are stored in the dictionary. And sometimes we are also interested in using some particular categories to store some special information in order to later enhance the bilingual dictionary. This is why ODLISTA was developed (on top of ODLISTER) which allows not only to list selectively but to create a file in another format.

Programs

The programming language in our system is PLNLP, a rule-based language with procedural capabilities, which can work in two modes: decoding (bottom-up parallel processing) and encoding (top-down sequential processing). The basic data structure is the record (set of attribute value pairs) and the typical objects dealt with by the language are graphs.

In our site, PLNLP is compiled into VM/LISP, and therefore the environment we are working on is LISP. However, the whole system should be portable into any other language for which a compiler for PLNLP exists, since we do not make use of any LISP feature.

The translator is therefore a LISP environment (SHLISPWS - shared LISP working space) customized for PLNLP needs. It should be emphasized that in this very same environment the English grammar (PEG), the translation system and the incipient Portuguese grammar are stored together.

The development of the system is easily achieved by having the rules being created in separate files in each developer's machine, rules which are then loaded into the LISP environment, tested, and modified till considered correct, by each person individually.

When several different new problems are considered to be solved, a new SHLISPWS is created by the organizer in charge, and carefully tested on several sets of sentences stored as measure of the system's performance. Only if there is no degradation is the previous workspace replaced, with some new sentences (and respective translations) added to the test corpus.

We plan to start the development of a program to help this checking task very soon.

Debugging

PLNLP provides several debugging facilities (such as various levels of trace) and almost unrestricted access to every detail of the system, being therefore easy to customize and use not only for programming translation rules but also for architecture design.

General utilities

We have a HELP for the Scientific Group, and plan to document there all facilities we develop, or that belong to the PLNLP / OD environment and are of general use.

As available tasks useful for a translation environment, the following can be mentioned:

- the program TRADUZIR, with the fact that it produces a list of words for which no translation was found, in alphabetical order,
- the listing of any dictionary and its possible conversion, ODLISTA,
- the merging of several updates to a main dictionary automatically, ODFUNDE,
- the adding of sets of changes to one dictionary instead of an entry by entry modification, ODBATCH,
- the PLNLP keep and test facility for parsing grammars, that allows a very efficient test over a significant number of sentences.

Several other functions that ease the development and especially the test of the translation system are planned to start soon:

- Automatic generation of corpora from raw text,
- Quasi-automatic verification of test corpora, to eliminate redundancy on a large set of test sentences,
- Improvement of the dictionary interface.
System organization

Machines

There are several virtual machines pertaining to the Group.

The one which userid is MENTOR has defined three permanent virtual disks with 191, 192 and 193 as virtual address respectively. They contain PEG and PLNLP files, as well as the dictionaries and several useful tools.

Each base worker is “owner” of a virtual machine. All of them have the same definitions. Their virtual machines are linked to disk 192 of MENTOR, which they can access in read/only mode.

The organizers working on this project can access all disks of MENTOR.

Two machines (called PLNLP and INTSYS) support CRITIQUE, an English text-critiquing program based on PEG and that is used by TRADUZIR.

MENTOR2 is the userid of a different virtual machine. Here some important files to the project are safe.

Finally, WSDIC stores WSMITH, an environment for human dictionaries, where we have available the Longman Dictionary of Contemporary English, the Webster’s 7th English dictionary, and ENGSPAN, from English to Spanish.

As a consequence of the situation depicted above, it should be clear that every modification done by a particular user in his/her own machine is not imposed upon other users. This means that everyone can have his/her own dictionary version, as well as computer scientists can have their own private rules. Until they are englobated in the translator system, they don’t interact with those of the other members.

Files

Among the programs installed at disk 192 of MENTOR (and therefore available to all workers on this project), we point out:
  - The PLNLP compiler
  - The translator programs, that is, the invoking EXECs (such as PORTUGA, PORT-RUN, ARV-RUN and TRADUZIR), and the executable version (PORTUGA SHLISPWS).
  - The dictionary access programs, the EXECs OD, ODT and ODG, and in general all dictionary utilities.
  - And finally, also the files containing the global test sentences are stored in that disk, so that everyone can test if her/his rules do not produce incorrect results in other occasions.

Among other programs installed on disk 193 of MENTOR (available to organizers only), we mention:
  - Exeecs CRIAPEG, CRIAPOR and CRIATRAD to create new environments.
  - PLNLP files and their respective compiled version (LISP files) for the complete system.
The amount of available documentation can be an accurate measure both of the work produced and of the importance given by its authors concerning its correct use and dissemination.

Our Group has produced mainly internal documentation, normally only broadcast to the other MENTOR teams and parallel NLP groups inside IBM, or Portuguese internal rules / work directives applying only inside the IBM-INESC Group.

There were however some documents intended for a broader audience, which are listed first:

  Used in several IBM sites as documentation for the PLNLP language.
  Master's dissertation.
- Common paper describing the work in MENTOR in Haifa, Madrid and Lisbon, not yet in print. The Portuguese contributions were co-authored by Regina Reis and Diana Santos.

As internal documentation, we have produced the following texts, which can be found in the collection MENTOR88/P one-year old, if their date is not later than October 1988,

- "Proposal of a PLNLP-based translation system", (DMS and PLM, Feb88)
- "Dealing with PEG's hidden pointer attributes in MENTOR88/P", (DMS, Feb88)
- "Lexical transfer in MENTOR88/P", (DMS, Feb88)
- "Report on a PLNLP-based translation system", (DMS and PLM, Mar88)
  This is a new (Mar88) version of the first paper, dated Feb88).
- "A set of procedures for transfer in MENTOR88/P", (DMS, Apr88)
- "The translation of infinitive clauses in MENTOR88/P", (DMS, May88)
- "Performance of the set of test sentences on INFCLs in MENTOR88/P", (appendix to the previous text)
- "Presentation of the work by the Portuguese branch of MENTOR", (DMS and PLM, Jun88)
  Copy of the foils presented at the MENTOR meeting in Lisbon.
- "Generation in MENTOR88/P", (PLM, Jul88)
- "Machine translation prototyping", (abstract for SJ describing the work in MENTOR on Haifa, Madrid and Lisbon).
- "Overview of work done in structural transfer", (DMS, Oct88)
Or they were joined in the collection Documentação do Grupo Científico IBM-INESC: primeiro semestre de 1989, where you can find:

- "Discussion on the several approaches for Lexical Transfer in MENTOR88" (DMS, Jan89)
- "Documentação sobre regras gerais de transferência" (DMS, May89)
- "Novo dicionário bilingue" (RR and DMS, May89)
- "Junção de novas regras ao sistema de tradução" (NA, May89)
- "ODLISTA: Manual do utilizador" (JCM, May89)
- "TRADUZIR EXEC: Manual do utilizador" (FP, May89)
- "ODBATCH: Manual do utilizador" (JCM, May89)
- "New bilingual dictionary" (RR, Jun89)
- "ODFUNDE: Manual do utilizador" (JCM, Jun89)

Not yet in a final form, there are already available the following pieces of documentation, in the same collection:

- "Tense transfer in PORTUGA" (RR)
- "Notes on PLNLP Grammar writing" (DMS)
- "TRADUZIR EXEC - Documentação Técnica" (FP)
- "Relatório: influência das conjunções no modo verbal" (JCA)
- "Relatório Adjetivos" (FG)
- "Relatório: Construções Reflexas" (AC)
- "Relatório NPs" (AC)
- "Verbos deféctivos, unipessoais e impessoais" (VF)
- "Relatório: Verbos preposicionais" (DR)
- "Ser ou não ser "it"" (VF)
- "Quantificadores" (RM)

In the above lines, DMS stands for Diana Santos, PLM for Paulo Libano Monteiro. RR for Regina Reis, NA for Nuno Alves, JCM for José Carlos Medeiros, AC for António Colaço, JCA for José Campos de Azevedo, DR for Dália Rosales, VF for Virgílio Fragoso, RM for Rui Marques, FG for Filomena Galvão and FP for Fátima Pedroso.
Present status and near future

This section is intended to describe the state of the art of our MT project, and what can be foreseeable in the near future.

Descriptions of the MT system will be put forward according to three different points of view:

- linguistic structures handled by the system.
- bilingual dictionary development.
- the system as a big software project.

This presentation will finish with the sets of test sentences we use for performance checking, plus a small example of a SCRIPT text in English that was translated globally.

It will be clear, that we do not claim that the system is complete in any respect, but we believe that, by continuing the work we started one and a half year ago, we can achieve relatively quickly a better performance.

Linguistic overview

Basically, all main structures are handled. We have developed more specific work on translation regarding the following subjects

- Infinitive clauses
- Arguments of a clause
- Tense transfer
- Negation
- Noun phrases' structure
- Contractions
- Lexical transfer

We expect to exhaust in one more year all syntactical structures of English, proceeding then with other tasks, like dictionary development, and multiword expressions.

Bilingual dictionary

Our Group was very fortunate to inherit an English to Portuguese dictionary from another machine translation project, which matched 16800 English entries into Brazilian Portuguese terms. That dictionary was adapted to our needs (converted to an OD form) and we used, apart from the words themselves, only the part of speech information.

The existence of that dictionary allowed us to concentrate on structural transfer, also making possible a broad experimentation both for that purpose and for the study of lexical transfer, before the considerable effort of building a dictionary was begun.

In May 1989, we decided to begin building our own dictionary. Several studies and experiments were conducted, before the actual building methodology was chosen, as well as its dimension and contents.
At this moment, the Group finds itself at the outset of a big effort, whose goal is to build an English-to-Portuguese dictionary that covers the "Vocabulário do Português Fundamental" (Fundamental Vocabulary for Portuguese), a list of 2070 words, plus around 100 words that do not belong there but appear in our corpora, plus those words that had already been edited and improved throughout our use of the system, whose number is around 200. This sums up to almost 2500 Portuguese words.

To get the English words that correspond to them, we used our previous dictionary, and got 3989 English entries with at least one translation belonging to our original set. It is at a dictionary of this size, namely around 4000 English words, (corresponding roughly to at most 6500 different Portuguese words), that we are thus aiming, in this first phase.

The building of this dictionary presupposes strict obedience to several rules and the use of an already existing human dictionary as the translations' source. Compatibility with PEG is also required, and the workers should introduce, for each entry, the conditions that are to be used for automatic disambiguation against the other senses of the word.

The fulfilling of syntactical features associated to words, on the contrary, will not be done during the creation of the entry, but will be a separate lexicographic work, in charge of the linguists who are performing and studying those specific tasks.

It is still too early to make an estimate of the time needed to build this dictionary, and the same applies to the percentage of correct translations that will be achieved by this method, and by our lexical transfer formalism.

Whatever time it will take, however, the results will definitely be relevant, not only to decide on the feasibility of such an approach, but as an effective evaluation of our lexical transfer formalism. In this respect, they will certainly point out the shortcomings and the more urgent needs, giving origin to a well-founded theory / formalism for that task.

It should also nevertheless be explained that the building of this new dictionary does not prevent that the old one continues to be used whenever an English word is not covered by the new one.

**System design**

The main architecture is already built, allowing for a correct processing of the whole system, giving access to alternative parses computed by PEG, and offering an interactive mode for lexical transfer.

There are still several improvements that should be attempted, though:

- Style processing, or how to accommodate several correct possibilities as translations of the same English sentence.
- Enhancement of generation capabilities (work in process).
- Improvement of the lexical transfer language.
- Development of tools to make it easy to test and measure the translation system.
- Improvement of the interactive disambiguation capabilities.

Even though these features are not required for a successful translation, they would definitely enhance the capabilities of the system as a powerful translation tool.

**Actual results**

General set of test sentences:

With all my friends I came last summer.
  \[=====>\] com todos os meus amigos eu vim verão passado.
All caught the last red train.
  \[=====>\] todos apanharam o último comboio vermelho.
Any discourses last.
  \[=====>\] quaisquer discursos demoram.
The process lasted many years.
  \[=====>\] o processo durou muitos anos.
The blue one stands there.
I do not like him.
I chose the blue ones.
She sings lovely old songs.
The yellow house has two little windows.
They want to eat apples.
I want to smoke.
He stopped to smoke.
She stopped the car to smoke a cigarette.
To swim is wonderful.
This is good to read.
It is good to read.
It is a good way to read.
The book describes how to swim.
They are assumed to come.
I have really to come.
I want not to smoke.
I want him to go.
I want the man to smoke.
I want the man not to smoke.
We are supposed to go.
They have really to come and see.
He chose between either to live or to die.
He is careful not to execute wrong commands.
He gets ready to return.
I want to read and write.
I want to read and to write.
I want him to read and to write.
I want him to read and write.
I have really to read and write.
I want to go to the market to see my mother.
eu quero ir ao mercado para ver a minha mãe.
This car is used to teach.
este carro é usado para leccionar.
I have the chance to learn and not to forget.
eu tenho a oportunidade de aprender e de não se esquecer.
The attempt to use the gun was wrong.
a tentativa de usar a espingarda era errada.
The instrument to use the gun is in the kitchen.
o instrumento para usar a espingarda está na cozinha.
The instrument to use is the gun.
o instrumento a usar é a espingarda.
I listened to the music to rest.
eu escutei a música para descansar.
I looked at the man to hear him.
eu olhei para o homem para o ouvir.
I expected all children to come.
eu esperei que todas as crianças vissem.
To want him to come is a violence.
quero que ele venha é uma violência.
I want him to play the violin and not to dance.
eu quero que ele toque o violino e não dance.
He tries to execute exactly the command you specify.
ele tenta executar exatamente o comando que tu especificas.
For you to determine it, do this.
para tu o determinas fazes isto.
the book is large enough to hold the story.
o livro é suficiente grande para conter a história.
it is large enough to hold.
suficiente grande para segurar.
He used him to steal chicken.
ele usou o para roubar galinha.
He used him to kill.
ele usou o para matar.
it is good for you to learn English.
és bom para ti aprender ingles.
this book is good for you to learn English.
este livro é bom para tu aprenderes ingles.
Every man came with hope.
todo homem veio com esperança.
Both children painted few fingers.
ambas as crianças pintaram poucos dedos.
Each book described how to find it.
cada livro descreveu como encontrar o.
Several parents like all children.
vários pais gostam de todas as crianças.
Many parents prefer small children.
muitos pais preferem crianças pequenas.
Small children came from both sides.
as crianças pequenas vieram de ambos os lados.
Two men had much hope.
dois homens tiveram muita esperança.
To want the man to pay to enter the house to buy cigarettes was to kill him.
quero que o homem pagasse para entrar na casa para comprar cigarros foi matar o.
That was the function to be used and not to show.
que ser a função para ser usada e para não mostrar.
it is OK to use the character string to do that.
ou é OK usar a cadeia de caracter para fazer isso.
the phrase grammar theory is used in natural language Processing.
a teoria de gramática de sintagma é usada em Processing natural de linguagem.
He isn’t a teacher.
ele não é um professor.
he has no money.
he doesn't like them.
he did like her.
he was never a good husband.
he couldn't stand it.
his wife never to hurt her.
the code quality is still not high.
the code quality is not always good.
there aren't any men in the cinema.
i see any trouble when it comes.
eu vejo qualquer dificuldade quando vem.
he is not very well.
he didn't quit not to hurt her.
ele não se demitiu para não a magoar.
he will arrive not to help her.
ele chegará para não a ajudar.
why didn't he come?
porque é que fez ele não veio?
when did she go?
quando é que fez ela foi?
who isn't here?
quem é que não está aqui?
will she come?
quererá ela virá?
he has been there yet.
ele não esteve lá contudo.
he will be there to help.
ele estará lá ajudar.
he'll be there.
ele estará lá.
he won't be there to help.
ele não estará lá ajudar.
the way commands are executed is this.
a mancira de os comandos serem executados é isto.
the man in the street is here.
o homem que eu encontrei na rua está aqui.
he is bigger than i am, but she is smaller than he is.
ele está maior do que eu estou, mas ela está mais pequena
do que ele está.
he is much more stupid than me.
ele é muito mais estúpido do que mim.
the value was returned by the function.
o valor foi devolvido pela função.
it is assumed that people can talk.
supõe se que as pessoas podem falar.
i like old men and women.
eu gosto de homens e mulheres antigos.
The last time I had some time to rest.
a vez passada que eu tive some marca para descansar.
I had some time to rest, last time.
eu tive algum tempo para descansar, a vez passada.
the gift was returned, but he never returned.
o presente foi devolvido, mas ele nunca voltou.
What are you looking at?
para o que é que estás tu a olhar?
I pointed out that he came.
eu apontei que ele veio.
I can sing if you want.

put the new tables and chairs in the room and clean everything.
puseste a nova mesas e cadeiras no sala e everything limpo.
Brig the new tables and chairs.

Functions are assumed to return one value only.

I shan’t thank you if you provide for that.

When i gave it to him he gave it to the old man.

Since he behaved strangely, she gave the young girl the book.

desde que ele se comportou estranhamente, ela deu o livro à rapariga jovem.

I saw the man who did it.

The book was given to him by George.

When i had already given it to her they came.

when the book was given to them it was too late.

i was not at the party, and they were unfortunately not in the garden

eu não estive na festa, e eles não estiveram infelizmente no jardim.

the ball you passed to him was blue.

a bola que tu lhe passaste era azul.

The meeting is in Lisbon.
a reunião é em Lisbon.

he is in Lisbon.
ele está em Lisboa.

He’ll be there at five o’clock.
ele estará lá em cinco o’clock.

he is not very well.
ele não está muito bem.

I am to do my homework.
eu devo fazer o meu homework.

to rest is to lose.
descansar é perder.

he is a teacher.
ele é um professor.

the painting was a work of art.
a pintura foi um trabalho de arte.

their coming was a nuisance.
a sua vinda foi um incômodo.

he is good at work.
ele é bom em trabalho.

he is always sad.
ele está sempre triste.

there are no problems.
não há problemas.

it is how he thinks.

why are you smiling?
porque é que estás tu a sorrir?

he was not alone because he was stupid.
ele esteve não só porque ele esteve estúpido.

he was not alone, because he was stupid.
ele não estava sózinho, porque ele estava estúpido.

when i wake i eat bread.
quando eu acordo eu como pão.

if he talks about me i sing a song.
A little text

Hello! This is a small text to show you how SCRIPT commands are processed.

We begin with an example.
This is the first sentence in the example.
You will see that this example is preserved, but not necessarily its "appearance" is.

Paragraphs in the examples are not respected, either.

Some examples with lists

Now we’ll see what happens to some lists.

1. This is the first item of an ordered list.
   It is very important that the list and list item commands are not in the same line. Otherwise
   your ordered list translation will begin with the second item.

2. If you nest several lists, you will end up with only one list in the end.
   See how that is done.
   a. This should be the first item of a new list.
      But, as you can see, the resulting translation has only one list, whose number of items is
      the sum of those of the two lists.
   b. Here is where the second list ends.

3. And this is the last item of the most exterior list.
   Some examples of definition lists are presented in the following lines. These lists are treated in a
   special way.

   definition nr. 1 This is the beginning of our definition list.

   definition 2 is the second one.

   It is sad, but these lists are not properly processed by our program. However, all
   terms are a sort of highlighted.

Citations and quotations

Quotations The same rule applies to inline quotations: We’ll show it to you with an universal
question. “to be or not to be, that is the question” should be translated?

And this is the end of our little text.

You certainly expected titles to have been properly translated. On the other hand, the words
highlighted in the translated text are words that were not found in the dictionary.

In spite of all its shortcomings, we expect TRADUZIR to be useful to you and we are happy to
provide it. Have a nice time!

And here is the translation:

Um pequeno texto

Hello! Isto é um texto pequeno para te mostrar como os comandos de SCRIPT são processados.
Nós começamos com um exemplo.

Isto é a primeira frase no exemplo.
Tu verás que este exemplo é preservado, mas a sua "aparência" não
necessariamente é.
Os parágrafos nos exemplos não são respeitados, também.
Alguns exemplos com listas

Nós veremos agora o que acontece a algumas listas.

1. Isto é o primeiro item dum lista ordenada. É muito importante que os comandos de item de lista e lista não estão na linha mesma. A tua tradução de lista ordenada começará de contrário com o segundo item.

2. Se tu encaixares várias listas, tu acabas com só uma lista no fim. Ve como isso é feito.

3. Isto deveu ser o primeiro item dum lista nova. Mas, assim como tu podes ver, a tradução resulting tem só uma lista, cujo o número de itens é a soma de aquele das duas listas.

4. Está aqui onde a segunda lista acaba.

5. E isto é o último item da lista mais exteriora.

Alguns exemplos de listas de definição são apresentados nas linhas seguintes. Estas listas são tratadas numa maneira especial.

nr de definição.
1.

Isto é o começo da nossa lista de definição.

Definição 2.

Seja o segundo. É triste, mas estas listas não correctamente são processadas pelo nosso programa. Porém, todos os termos são um espécie de highlighted.

Citações e citações

Citações A regra mesma aplica a citações de inline: Nós mostraremos te com uma pergunta universal o. " para ser ou para não ser, aquele seja a pergunta σ dever ser traduzido?

E isto é o fim do nosso pequeno texto.

Tu esperaste certamente que os títulos tivessem sido correctamente traduzidos. Na outra mão, as palavras realçadas no texto traduzido são palavras que não foram encontrados no dicionário.

Apesar de todas as suas deficiências, nós esperamos que o TRADUZIR seja útil a ti e nós estamos felizes por o providenciar. Te um tempo simpático!
External opinions

We asked some kind of evaluation of our work to people who know it and whose opinion is certainly relevant. We are happy to be able to present the following answers, from names of outstanding researchers in NLP inside IBM.

MORI RIMON was the founder of the MENTOR project, and the head of the Israeli team for this project.

Date: 30 May 89, 12:50:38 IDT
From: Mori Rimon +972-4-296-207 (211) RIMON at HAIFASC3
Department Manager, Software Systems and Algorithms
IBM Israel Science and Technology, Haifa, Israel
To: Diana M.S.M.P. Santos PODIANA at LISHQ
cc: Andre Emonet PO34814 at LISHQ

Re: Your notes of 22 and 23 May 1989

Dear Diana, I'd like to take this opportunity to summarize my view on your contribution to the multi-national MENTOR project on Machine Assisted Translation, and especially on your prototype MENTOR88/P.

MENTOR was started in 1987 as a research project, the objective of which was to enhance IBM's understanding of Machine Assisted Translation. At that time, there was no other significant multi-target MAT project in IBM and, generally, not enough work was underway on Natural Languages other than the major. Since MAT problems are very difficult, and one must have short range milestones even in a long term project, we have decided to develop preliminary prototypes in a one year time frame. We first hoped to only have two such prototypes available within this short time: English/Spanish and English/Hebrew. It was a great surprise to realize that the small group in Lisbon is making very nice progress and can also meet the mid 88 deadline with a running English/Portuguese prototype (the fourth MENTOR group, in Finland, did not deliver a prototype but has been rather developing a proposal for an alternative approach based on Unification formalisms).

The Portuguese prototype (like the other two), while limited, still features all translation phases. I.e., it can take English sentences all the way through analysis, bilingual transfer, and target language sentence generation. A unique aspect of the Portuguese prototype is its adherence to Yorkton's PEG's philosophy and the usage of PLNLP alone for implementation. Since the main idea of the prototyping stage was to experiment with different approaches, the Portuguese scheme was definitely valuable in providing deep understanding of the power and shortcomings of a PEG/PLNLP-based system. Let me add that the lessons learnt by developing all MENTOR prototypes contributed significantly to the evolvement of the current ideas about IBM MAT products and to the very recognition of MAT as a worthy and feasible business opportunity.

Among the other important technical achievements of the Portuguese prototype, I'd like to mention the following:
- The output of the Transfer phase is a PEG-like graph for Portuguese.
- The graph is traversed just once in a structured way in both
Transfer and Generation.
- Lexical Transfer and Structural Transfer are interleaved.
- Linguistic knowledge is stated in declarative terms.
- Interaction is allowed with a user to select a lexical mapping.
- Generation is based on sparse dictionaries techniques.
- The Morphology is rule-based.

Personally, as the project leader it was a pleasure for me to work with you and your colleagues through the first stage of MENTOR. You have certainly achieved a lot with a very small number of people. I hope that you will continue to contribute both to local NLP efforts in Portugal and to any MAT Research & Development efforts IBM will decide to undertake in the future.

Mori Rimon, Haifa

PAULA NEWMAN is the responsible for the Los Angeles Scientific Center project in Machine Translation, and one of the responsible for Computer Assisted Natural Language Translation in the IBM Europe Institute 1989.

Date: 23 May 1989, 08:28:42 PDT
From: Paula Newman 8-882-5423 PAULA at LOSANGEL
To: Diana Santos PODIANA at LISHQ
cc: Andre Emonet PO34814 at LISHQ

Subject: MENTOR-Portuguese

Diana,

The work described in your paper, as usual, reflects a pragmatic approach in the best spirit of the term - doing as much as is possible with the "front-end" adopted for MENTOR, in a well-designed way, with painstaking attention to the actual contrastive facts, and impressive productivity.

While at LASC we are trying to obtain an architecture suited to m-to-n systems, it is likely that the IBM MAT direction for the immediate future will be 1-to-n. Your overall approach, if combined with a source analysis phase assuming its proper responsibilities for input text disambiguation, is one which can produce an excellent result at acceptable cost.

Among the important architectural ingredients exemplified by your prototype are: (a) a top-down transfer process giving the opportunity for major restructuring when necessary, (b) a significant independent generation phase, limiting transfer, and (c) careful attention to succinct lexicon specification - one of the keys to cost-effective translator development.

Regards,
Paula

KAREN JENSEN is the author of PEG, the English grammar that we use in our machine translation system. She is also the mentor of the Portuguese grammar.

She belongs to IBM Research, Yorktown Heights, and is in a two-years assignment in ASD Bethesda, for development of CRITIQUE, a text-critique system which uses PEG and for which she has recently received an Outstanding Innovation Award.

Date: 2 June 89, 20:46:28 SET
From: KJENSEN at PARIS
To: PO34814 at LISHQ
cc: PODIANA at LISHQ
Dear M. Emonet,

Diana Santos has asked me to write a brief evaluation of her work on organizing and implementing the Portuguese PLNLP grammar. I am happy to do this.

Diana has chosen to develop the grammar by dividing the work among several people, including both linguists and computer scientists. This approach to grammar development is accepted in the profession, and has been taken by several well-known natural language processing groups. I am reminded, in particular, of the work done in California by, for example, Hewitt-Packard on their HPSG grammar of English.

Within this framework, Diana's distribution of tasks seems sensible. As she has explained it to me, linguistic analysis of Portuguese, which is now being done for the MT project, will be equally applicable to the Portuguese parsing grammar. In this category, the following sorts of problems are being studied: reflexive constructions, tense and aspect, defective verbs, and subcategorization of adjectives and conjunctions. All of this work should be directly applicable to the parsing grammar. In fact, there is a distinct advantage in a work environment that combines the development of the Portuguese grammar with the development of an MT system: it is much easier to guide the grammar writing when it can be oriented toward the needs of an independent application like MT.

Currently, Diana alone is responsible for writing the Portuguese grammar. The next step in her development plans, as I understand, is gradually to give her co-workers responsibility for the actual grammar writing. They have been trained, according to her, in our system and in the interpretation of our English grammar. Once she has other grammarians who can handle the actual rule writing, then she herself will be free to devote more of her time to other projects (such as her Ph.D. dissertation research).

The PLNLP grammar of Portuguese, like most other PLNLP grammars now being written, is in its beginning stage. Nevertheless, its scope and coverage are comparable to, and sometimes greater than, many of these other grammars. There are two main reasons for this state of affairs:

1. Other grammarians have been required to work on dictionary matters for much of the time since they left their Bethesda training. Diana has, for her part, been required to spend much time on the Portuguese MT system, as you know. So the time available for actual grammar development tends to even out, in many (not all) countries.

2. Diana herself is an exceptionally quick worker. Because of her talents for language analysis and for programming, she has been able to accomplish an amazing amount in a short time.

It is certainly possible to develop a computational grammar within the framework that Diana has laid down, where the responsibility for the analysis of the language would be shared and discussed among various technically competent colleagues. Given Diana's organizational and other abilities, her energy, and her enthusiasm for the project, I would expect that Portugal could play an internationally significant role in Natural Language Processing within IBM, in a rather short time.

Sincerely,

Karen Jensen

GEORGE HEIDORN is the creator of the PLNLP system and one of the most outstanding NLP researchers both inside IBM and outside.
He belongs to IBM T.J. Watson Research Center, and is currently on assignment in ASD Bethesda, in order to develop PLNLP further as the most suitable environment for multiple natural languages processing.

He was the first editor of the well-known magazine “Computational Linguistics”.

Date: 8 June 89, 19:19:35 ET
From: HEIDORN at BETVTVM1
To: PO34814 at LISHQ
cc: PODIANA at LISHQ

Dear M. Emonet:

I would like to say a few words in support of Diana Santos and her Research Group. The work that is being done there is top-notch.

Natural language processing is an area whose importance is becoming increasingly well known all over the world, both inside IBM and outside IBM. The first computer company that supports many human languages in a consistent manner is going to have a tremendous competitive advantage; and we want IBM to be that company. Portugal may be a small country, but nevertheless the Portuguese language has an important place among the languages of the world and should not be left out.

The Group in Lisbon is poised to make a significant contribution. The work that has already been accomplished is quite remarkable. Many original ideas have come from there, and I expect many more.

Combining the work on machine translation and grammar writing in the manner that Diana has done will improve both parts of the work and serves as an excellent model for similar work in other countries. Documentation produced by Diana for training her colleagues at INESC is very well done, and is also being used at many other IBM sites around the world. Diana has been very active in assisting her colleagues at PLNLP workshops, and has suggested many things to me that have improved our system. She is both a capable researcher and a practical engineer, in addition to being an excellent organizer.

I realize that I do not know all of the fine points of IBM-Portugal business, and that you have many considerations to take into account in deciding the fate of this Group. But I can tell you that you have a unique situation there, and I hope that you will be able to give it special treatment. IBM needs this work to continue.

Sincerely,

George E. Heidorn
Manager, Natural Language Processing, IBM Research
(On 2-year assignment to ASD-Bethesda)
Conclusions

In our opinion, the IBM-INESC Scientific Group is the ideal place to develop this machine translation effort for Portuguese:

- There is nowadays a multidisciplinary group with working knowledge of the problem and of our solutions.
- There are no other places inside IBM where Portuguese is being dealt with for machine translation. And our work has been praised by other NLP groups inside IBM.
- There is at the same time a PLNLP grammar being written for Portuguese, which is optimally integrated in the MT environment and can very fruitfully interact, both to MT and grammar development interests.
- It is advantageous for IBM Portugal to have a good relationship with INESC and therefore contribute to research and development in this country. An institution as INESC is, moreover, the optimal place to do research due to its academic environment and its privileged relationship to University.
- This Group does not represent a heavy monetary effort for IBM Portugal, as it can be mostly financed by the funding for the Portuguese grammar, coming from abroad.
- Interaction with NLS of IBM Portugal, which hires nowadays more than fifty translators to satisfy the need for product documentation in Portuguese, could mean for IBM a significant improvement in economy and consistency in a future not very far, while would at once solve the bilingual dictionary development requirements of our machine translation system.