

The GRACE French Part-of-Speech Tagging Evaluation Task

Gilles Adda Joseph Mariani Josette Lecomte Patrick Paroubek Martin Rajman

LIMSI, France
gadda@limsi.fr

LIMSI, France
mariani@limsi.fr

INaLF, France
josette.lecomte@inalf.cnrs.fr

LIMSI, France
pap@limsi.fr

EPFL, Switzerland
martin.rajman@epfl.ch

Abstract

The GRACE evaluation program aims at applying the Evaluation Paradigm to the evaluation of Part-of-Speech taggers for French. An interesting by-product of GRACE is the production of validated language resources necessary for the evaluation. After a brief recall of the origins and the nature of the Evaluation Paradigm, we show how it relates to other national and international initiatives. We then present the now ending GRACE evaluation campaign and describe its four main components (corpus building, tagging procedure, lexicon building, evaluation procedure), as well as its internal organization.

1. The Evaluation Paradigm

The Evaluation Paradigm has been proposed as a mean to foster development in research and technology in the field of language engineering. Up to now, it has been mostly used in the United States in the framework of the ARPA and NIST projects on automatic processing of spoken and written language.

The paradigm is based on a two step process:

- first, create textual or voice data in the form of raw corpora, tagged corpora or lexica, which are then distributed to main actors in the field of language engineering for the realization of natural language processing tools. These tools address problems like disambiguation, natural language database query, message understanding, automatic translation, dictation, dialog, character recognition, etc.;
- then, test and compare systems on similar data. The results of the tests and the discussions (within specific workshops, for example) triggered by their publication and comparison constitute a good basis for the evaluation of the pros and cons of the various methods. The resulting synergy is a dynamizing factor for the field of Language Engineering. The Linguistic Data Consortium (whose function is to collect language related data and to organize their distribution) is a typical illustration of positive consequences of programs implementing the Evaluation Paradigm.

The GRACE evaluation program is meant to be an implementation of the Evaluation Paradigm in the field of morpho-syntactic tagging. As such it corresponds to an evaluation campaign of Part-Of-Speech taggers for French organized within an automated quantitative black-box evaluation framework.

2. The GRACE evaluation program

Started upon the initiative of Joseph Mariani (LIMSI) and Robert Martin (INaLF), GRACE (Grammars and Resources for Analyzers of Corpora and their Evaluation) was part of the French program: "Cognition, Communication Intelligente et Ingénierie des Langues" (Cognition, Intelligent Communication and Language Engineering), jointly promoted by the Engineering Sciences and Human Sciences departments of the CNRS (National Center for Scientific Research).

The GRACE program was intended to run over a four year period (1994-1997) and was planed in two phases: a first phase dedicated to Part-of-Speech taggers, and a second phase concerned with work on syntactic analyzers, but which has been abandoned.

The first year was devoted to the setting up of a coordination committee in charge of running the project and a reflection committee.

The responsibility of the reflection committee formed with a pannel of experts from various domains, is to define the evaluation protocol, to specify the reference tagset and lexicon, to decide which data will be made available to the participants, and to organize the workshop for the presentation of the final results.

The third entity of the GRACE organization regroups all the participants. They come both from public institutions and industrial corporations. Only participants with fully operational systems were allowed to take part in the evaluation. Furthermore, only the participants which have agreed to describe how their system works (at least during a workshop whose attendance would be restricted to the sole participants) were authorized to take part in the workshop concluding the evaluation campaign.

20 participants, from both academia and industry, registered at the beginning of the project. During the project, this number slightly decreased. 17 took part in the dry-run and 13 in the final test the results of which will be published at the beginning of fall '98.

3. Defining the Evaluation Procedure

For the definition and the organization of the GRACE evaluation campaign, we build upon the work done in previous evaluation programs, in particular the evaluation campaigns which have been conducted in the United States, especially in the scope of ARPA Human Language Technology program. Namely:

- the MUC (MUC-1, MUC-2, MUC-3 (Sundheim 1991), MUC-4 (MUC 1992)) conferences, aiming at the evaluation of message understanding systems¹;
- TIPSTER, concerning the evaluation of automated information extraction systems from raw text data;
- the TREC (Harman 1993; Harman 1994) conferences, concerning the evaluation Information Retrieval systems operating on textual databases;
- ParsEval and SemEval, which find their origin in Ezra Black's work (Black 1991; Black 1993; Black 1994) on the evaluation of syntactic analyzers done within the scope of an ACL working group.

GRACE also looked at the “Morpholympics” competition (Hauser 1994a; Hauser 1994b), which was organized in spring 1994 at Erlangen University in Germany for the evaluation of German morphological analyzers.

MUC and TREC use *task oriented black-box* evaluation schemes requiring no knowledge of the internal processes or theoretical underpinning of the systems being tested, while ParsEval and SemEval (some of which will be part of MUC-6) are approaches which attempt to evaluate systems at the module level by using a benchmark method based on a reference corpus annotated with syntactic structures agreed upon by a panel of experts.

An additional list of evaluation methods for linguistic software (*lingware*) now in use in the industry was found in Marc Cavazza's report (in French) for the French Ministry of Education and Research (Cavazza 1994). An other extensive overview of evaluation programs for Natural Language Processing systems is provided in (Sparck Jones and Galliers 1996).

Similarly to the evaluation campaigns organized in the United-States, GRACE was divided into four phases:

1. training phase (“phase d’entraînement”): distribution of the training data (the *training corpus*) to the participants for the initial set up of their systems;
2. dry-run phase (“phase d’essais”): distribution of a first set of data (the *dry-run corpus*) to the participants for a first real-size test of the evaluation protocol;

¹

the task used in the MUC evaluation campaigns was for the systems to fill in predefined frames from texts relating US Navy manoeuvres (MUC-1 and MUC-2) or terrorism acts (MUC-3 and MUC-4)

3. test phase (“phase de test”): distribution of the “real” evaluation data (the *test corpus*) to the participants and realization of the evaluation;
4. adjudication phase (“phase d’adjudication”): validation with the participants of the results of the evaluation; this phase leads to the organization of a workshop where all the participants present their methods and their systems and discuss the results of the evaluation.

According to the task-oriented approach chosen in GRACE, the evaluation procedure was based on a automated comparison, on a common corpus of literary and journalistic texts, of the PoS produced by various tagging systems against PoS manually validated by an expert (tagging is therefore the task selected for evaluation).

In addition, as the evaluation procedure has to be applicable for the simultaneous evaluation of several systems (that may very well use various tagging procedures –statistical, rule-based, ...), the definition of the evaluation metrics cannot rely on any presupposition about the internal characteristics of the tagging methods. It has therefore to be exclusively defined in terms of the outputs produced by the systems (pure “black-box” approach), which, in the case of tagging, can be minimally represented as sequences of pairs the elements of which are the word token and its tag (or tag list).

Such an output is considered to be “minimal” for the tagging task because several taggers also produce some additional information in addition to the simple tags (e.g. lemmas). In GRACE, we decided to not take into account such additional information (for example, no evaluation of the eventual lemmatization provided by the systems was performed) and restrict ourselves to the tagging task, defined as aiming at associating one unique tag to each token (and not for instance a partially disambiguated list of tags, which would have required a much more complex metrics for comparing the systems).

However, even with such a minimalistic definition for the tagging task, the actual definition of a working evaluation metrics did require from the GRACE steering committee to take several decision about various crucial issues:

- how to compare systems that do not operate on the same tokens (i.e. use different segmentation procedures)? How to take into account the processing of compound forms?
- how to compare systems that do not use the same tagsets ?
- how to weighten in the evaluation the different components that build up any realistic tagging system? In particular, how to evaluate the influence of the capacity of a tagger to handle unknown words ? How to evaluate the influence of the quality of the lexical information available?

Build upon the evaluation scheme initially proposed by Martin Rajman in (Adda et al.(1995)) and then adapted and

extended by the GRACE committees, the evaluation procedure used in GRACE is characterized by the following aspects:

Dealing with varying tokenizations The problem of the differences in the variations of text segmentation between the hand-tagged reference material and the text returned by the participants is a central issue for tagger evaluation. Indeed, Not to leave a complete freedom to the participant about the tokenizing algorithm (and the lexicon) used to segment the data, they had to tag, seemed us unrealistic. As a consequence to compare the text tagged by a participant with the reference text which has been hand-marked we need to identify the tokens of both texts in order to be able to compare the tags attached to them.

We solved this problem defining a re-alignment procedure wrapped around the UNIX command *diff* (with its -D option) applied to streams of characters (produced by “exploding” the words into sequence of characters, e.g. one per line) in conjunction with a pivotal tokenizing scheme of the finest possible grain needed for our purpose. Reference tokens are always of a size smaller or equal to the one of the participant’s token they are re-aligned with, because we consider a form to be any sequence of characters of a class different from the separator class, defined as containing the white space, the newline, and all punctuation characters, inclusive of the hyphen and the apostrophe. If some tokens are of a grain bigger than the one of the reference segmentation scheme, they are re-tokenized in a pre-processing step, i.e. before being transformed into a stream of characters.

The result yielded by *diff* is then handled by two subsequent procedures, the first one generates “ghost” characters in order to produce two sequences of characters of exactly the same size, the second rebuilds the original tokens from the now aligned sequences of characters and make explicit the alignment information while attaching anew the tags to their respective forms and filtering out the data that could not be re-aligned properly (on average 0.5%).

The procedure that reconstructs the original tokens distinguishes two types of areas in its input data, places *diff* has recognized a straightforward alignment, and places where ghost characters had to be introduced in order to resynchronize the two character streams. We call these “fuzzy zones” because they can contain forms for which no realignment can be found.

In a first step, the reconstruction procedure will propose candidates matchings between the tokens contained in the two streams. Note that these matchings are not necessarily of the one-to-one kind, but can also be of the one-to-many kind.

Until now, the best results have been obtained with a double exploration of the fuzzy zones, first by looking for strict string-equality in order to identify a match between participant and reference tokens, then, starting from the beginning of the zone, by accepting for the yet unmatched tokens, the first case encountered where one token is a substring of the other (without case distinction) as a valid match.

Because it remains generic, this algorithm leaves open a large set of possibilities for refining the alignment, in par-

ticular for resolving the one-to-many possible matches with ad-hoc patterns bound to specific linguistic phenomena or formatting practices. A few heuristics of this kind have been implemented to take care of the idiosyncrasies held in the data of a few participants (e.g. systematic expansion of the contracted article **du** into **de le**).

The token loss rate over the portion of the test corpus used for the measurements (103,676 tokens), averaged over all the participants to the tests and over both the participants and the reference tokens is: 0.575%, with extremes of: 0.029% and 1.89%.

In addition to a good success rate, this method presents the advantage of identifying explicitly the forms which could not be re-aligned and thus giving the means for their automatic elimination.

The performance measures are being computed only on the data which have been properly re-aligned with the reference data; evaluation is turned-off on the portions of text that the algorithm cannot re-align properly.

In order to complete our work on the re-alignment procedure, we would need to formalize this algorithm, and in particular to compare it with the one implemented in the GNU utility *wdiff* (*diff* for the words), to precise why our first experimentations made us discard it.

Taking into account several tagsets As it was of course not possible to impose on the participants to use the same unique tagset, it was decided in GRACE to first define a *reference tagset* (also called hereafter the GRACE tagset) well adapted for the morpho-syntactic description of French and then to ask the participants to provide a *mapping table* allowing to project the tags used in their system into the GRACE tagset.

In order to make such kind of mapping possible, a necessary condition was that the reference tagset is the finest among the tagsets defined by the participants (or that the participants accept not to represent in the tags that their system produces the information that could not be taken into account in the reference tagset).

The reference tagset: For the definition of the GRACE tagset, we were inspired by the work done at the University of Pennsylvania for the tagging of the Penn TreeBank and by the tagging recommendations issued by European projects such as MULTEXT and EAGLES. The GRACE tagset is therefore directly derived from the MULTEXT/EAGLES standard (Leech and Wilson 1994; Veronis and Ide 1994), yet with some differences concerning some PoS or attribute tags.

The morpho-syntactic descriptions were broken down in Part of Speech (main category) and list of attributes, thus defining detailed morpho-syntactic patterns.

The tagset actually evolved during the three phases of the GRACE action (training, dry-run and test): modifications were raised from experience in tagging and all participants were invited to suggest and discuss modifications. Some of them were very active, others never reacted.

The final GRACE tagset is therefore the result of a consensus between the GRACE reflection committee and the participants, obtained by means of a circular step-wises re-

inement procedure, started upon a proposition of the organizers. It contains 12 main categories (9 for Parts of Speech, 1 for punctuation, 1 generic class, and an evaluation blocker). Altogether, the tagset comprises 311 different full tags.

Notice that we did not use the techniques and tools described in (Teufel 1995) to derive the GRACE tagset from the comparative analysis of the tagsets of all the participants. The reason for this was that it took a certain time to setup a contractual environment in which all the participants accepted to provide the description of their tagsets (the tagset description corresponds indeed to a crucial, and therefore valuable, part of a tagger). A description of all the different tagsets was not available early enough to allow the use of comparative techniques.

In addition to the list of tags, the GRACE organizers established for each PoS a list of authorized "patterns" (i.e. feature combinations corresponding to linguistically sound descriptions). This was of great help for the building and validation of the mapping tables, as it allowed the creation of automatic verification tools.

Furthermore, a document describing tagging guidelines was also produced and was quite useful for those of the participants who were interested in the way the reference texts had been tagged. A first draft of the guideline was actually distributed to all the participants, but was not further updated, as it appeared not to be a pre-requisite for starting to work on the corpora. However, the tagging guidelines were discussed throughout the validation part of the dry-run. Needless to say it was really useful to the human expert who was in charge of preparing the reference tagged corpus, and performed minimal consistency checking.

Finally, to illustrate the choices taken and facilitate the discussion, a small corpus of 200 sentences made of both artificially constructed sentences and excerpts extracted from real corpora was produced, hand-tagged and communicated to all the participants before the dry-run phase.

Mapping tables: As the reference tagset was defined as the finest one among the different tagsets used by the participants, the translation from one tagset into the common tagset is potentially a many to one correspondence (ideally a one to one). An example of a excerpt from a mapping table is given in figure 1.

PUL →	X
REL →	Prms Prfs Prmp Prfp
SBC →	Ncms Ncfs Ncmp Ncfp

Figure 1: Excerpt from a mapping table

Out of the initial 21 registered participant, 14 have provided a mapping table. Notice that a few participants (GR-EYC - J. Vergnes et INGENIA - P. Constant) could not directly provide a mapping table because in the framework of their systems (being actually parsers, specialized, for the GRACE evaluation, to the tagging task) the extraction of the information required to perform the mapping would necessitate a too large effort because of the specificity of the linguistic formalism they used (syntactic links). For the dry-run phase, these participants did perform themselves

the mapping of their internal "categories" directly into the GRACE tagset.

Inverse mapping: As it is not always possible to have a one-to-one mapping between the tags of two formalisms while building a mapping table, it would be better to use two tables (as remarked G. Lallich-Boidin and M. Bertier of the Cristal team), one for the direct map and the other for the inverse map. Because it was not practical to the participant to craft themselves this inverse mapping table, we decided to build it from the direct map using an automatic procedure, at the price of an extra amount of ambiguity produced by the computation of the inverse mapping. This procedure has been tested on the dry-run corpus of the INaLF and Limsi (both using E. Brill tagger trained on different corpora). The validity of this approach still needs to be validated on real-size data for all the other participants but the first tries have proved to be encouraging. Given the two input parameters made of:

- 1) a mapping table:

Participant tag (rule head)	GRACE tag (rule body)
tagP0 →	tagG00 tagG02 tagG03 ...
tagP1 →	tagG10 tagG12 tagG13 ...
.	...

- 2) a GRACE tag list: LG0 | LG1 | LG2

the function computing the inverse mapping will try to find one or several combination of the participant tags which could plausibly have generated the tag list. At each stage of the computation, to all the rules (corresponding to entries in the direct mapping table) which were selected during the previous stage, the following filters will be sequentially applied to select the list of rule whose left-hand side member (participant tags) will constitute the desired results:

- stage **0**, input of the direct mapping table and of the list of tags to analyze (second input argument).
- stage **1**, selection of all the rules.
- stage **2**, selection of all the rule combinations whose union of the tags contained in their right-hand side maximally covers the input tag list (second argument).
- stage **3**, selection of all the rules which maximize the number of tags belonging to their right-hand sides also present in the input list (second argument).
- stage **4**, selection of all the rules which minimize the number of the tags contained in their right-hand sides missing in the input list (second argument).
- stage **5**, selection of the rule combinations which minimize the number of tags which overlap across the different right hand sides (minimal intersection of the right-hand sides of the rules associated to the selected tag combination).

- stage **6**, display of the head of the rules (left hand sides, i.e. participant tags) belonging in the rule combination which has got the best score at the end of stage 5.

For instance if the input is,

E1 →	C1 C2 C3
E2 →	C2 C4
E3 →	C1 C3

with the second argument: C1 | C2 | C3 | C4,
the result will be E1 | E2

4. Defining the Evaluation Metrics

Precision and decision In GRACE, evaluation is done by analyzing the output produced by the systems. For this purpose, two specific quantities, "Precision" and "Decision", inspired from the "Precision" and "Recall" used for the evaluation of information retrieval systems, were defined.

Precision measures the proportion of correctly tagged tokens within the set of all the tokens that were non ambiguously tagged by the evaluated system. It is therefore a measure of the accuracy of the tagging effectively performed by the system.

Decision measures the proportion of tokens non ambiguously tagged within the set of all token processed by the evaluated system. It therefore quantifies to which extent the evaluated system effectively tags the input data.

More formally, for each token identified in the input data, a tagging system can:

- associate with the token a single tag compatible with the reference tag that was associated with the token in the reference corpus (more precisely, compatible with the tag that was associated with the token in the reference corpus mapped with the token identified by the system). In this case, we say that the system made a *correct tagging*.
- associate with the token a single tag not compatible with the reference tag that was associated with the token in the reference corpus. In this case, we say that the system made an *incorrect tagging* (also called an *error*).
- associate with the token a set of tags, none of which is compatible with the reference tag that was associated with the token in the reference corpus. This case is another possible instance of an *incorrect tagging*.
- associate with the token a set of tags, among which at least one is compatible with the reference tag that was associated with the token in the reference corpus. In this case, we say that the system made an *silence*.

In addition, a token can be excluded from the evaluation (if for instance it appears in a zone where the token re-alignment could not be properly performed). In this case, we say the we have a *non evaluation*.

If we then note :

Nb	total nb. of tokens
NbOK	nb. of correct taggings
NbERR	nb. of errors
NbSIL	nb. of silences
NbNonEval	nb. of non evaluations

we can define the precision P and the decision D as:

$$P = \text{NbOK} / (\text{NbOK} + \text{NbERR})$$

$$D = (\text{NbOK} + \text{NbERR}) / (\text{Nb} - \text{NbNonEval})$$

The evaluation of a tagging system can then be characterized by the pair: (P,D), and system comparison is achieved by measuring the distance between the (P,D) positions associated to each system in the area defined by $[0,1] \times [0,1]$. For a detailed presentation of GRACE measures, see (Paroubek 1997) (in French).

The finalization of the definition of the evaluation metrics used for the test phase were done in interaction with the participants at the end of the dry-run phase.

In GRACE, all the systems are evaluated on the same data, but the existence of mapping functions to and from (see previous paragraph) the reference tagset and the various tagsets used by the different participants enables several kinds of performance measures:

- "absolute performance measure", i.e. relatively to the reference lexical descriptions themselves;
- "relative performance measure", i.e. relatively to the tagset used by the system itself,
- "clustered performance measure", i.e. relatively to the tagset of another system, within classes (i.e. subsets of systems), according to a set of applications classes defined in collaboration with the participants. Such an approach thus provides some means to take into account possible biases induced by a particular application domain.

5. Building the Corpora

Text corpora of sufficient size are required for corpora based systems (see summer ESCA-Elsnet "Corpus Based Methods" in July 1994).

According to the EAGLES report on NLP-system evaluation (EAGLES 1994), corpora have proved to be useful mostly in *adequacy evaluation*, and *progress evaluation* leaving aside *diagnostic evaluation*, which for instance was more the concern of projects like TSNLP.

Generic recommendations on the properties re-usable evaluation data ought to display are proposed in (Crouch et al.1995), the report of the study group on evaluation, set up within the EAGLES framework to specify guidelines for assessment and validation of LE projects in the Fourth Framework Program. According to this source, evaluation data must be:

- *realistic*, i.e. be of the same kind as the data received by the system or component being tested during its normal operation,

- *representative*, i.e. contain instances from the full range of input data that would be normally received by the system or component under evaluation,
- *legitimate*, i.e. easy to acquire, or widely reusable for other purposes

(Crouch et al.1995) identify two important properties of the evaluation which strongly condition the requirements put upon test data:

- the *granularity* at which systems will be evaluated (e.g. at the level of user-significant tasks only, or at some level of tasks that are user-transparent),
- the *generality* at which systems will be evaluated (e.g. how much do the linguistic features of the provided input data (resp. expected output date) vary, relatively to the characteristics of the data in the intended application? is the evaluation done against data from different languages, different domains, etc.?).

The authors also remark that in most cases, large corpora do not suffice by themselves as the basis for an automatic evaluation procedure. Such corpora need to be annotated depending on the system being tested. But in this case, most annotation schemes are specific to a particular class of application, and therefore hardly re-usable. To solve this problem and other difficulties related to evaluation, Klaus Netter has proposed to use, in conjunction with glass-box evaluation², *layered annotations* for corpora. This can be done at different levels of abstraction matching the intermediate reference levels of the considered evaluation scheme. The annotations could include all kinds of information, such as morpho-syntactic tagging, word sense disambiguation, phrase structures, relational structures, semantic representations including resolved references as well as annotations specific to the application being evaluated.

Since there is no established corpus of written text for academic work in France, only two different genres were available, literary and journalistic. Two sources were used to establish corpora: the FRANTEXT text corpus (INaLF) mainly consisting of literary texts from the XIXth century or the beginning of the XXth century (about 160 million words), and the French newspaper "Le Monde" for which corpora are regularly distributed on a CDROM (initially 50 millions words were available, but several hundred million of words are now available with the publication of the archives of the newspaper dating back from 1987).

We have solved the legal issue of copyrights in two ways:

- by selecting from the FRANTEXT database texts which have no copyright restrictions imposed on them.

²

(EAGLES 1994) defines *glass-box* evaluation as requiring knowledge of the internal working and theoretical underpinning of the system being tested, while black-box evaluation only sees the final output and its relationship to the original input. *Black-box* evaluation is typical of *adequacy evaluation* of market products.

- obtaining from the "Le Monde" newspaper, through the LIMSI laboratory, the authorization to use and distribute their material, under the condition that each recipient would sign an agreement forbidding redistribution and commercial use of the data it receives.

The data has been separated into three packages each corresponding to a phase of GRACE campaign, all with a roughly balanced distribution of texts between the two sources (the "Le Monde" corpus and the FRANTEXT database):

- a training corpus of 9 million word forms, with roughly 4 million word forms out of "Le Monde" corpus and 5 millions out of FRANTEXT.
- a dry-run corpus of 450,000 word forms, split between a sample of 110,000 out of "Le Monde" and another of 340,000 word forms from FRANTEXT. A portion of around 26,000 words (13,000 FRANTEXT and 13,000 "Le Monde") has been hand-tagged to serve as the evaluation corpus;
- a test corpus of 650,000 word forms, with roughly 300,000 coming from "Le Monde" and 350,000 coming from FRANTEXT. About 110,000 (65,000 for FRANTEXT and 47,000 for "Le Monde") have been hand-tagged.

The texts extracted from the FRANTEX database were essentially novels, along a few travel accounts, bibliographies and other such pieces of literature, all written in prose (avoiding theater and poetry) at the end of the XIXth century or the beginning of the XXth. The texts from the "Le Monde" newspaper were randomly selected from recent editions of the newspaper. The resulting data has been drowned inside additional data (taken from the over 4,000 texts of the FRANTEXT database and from the available "Le Monde" CD-ROM) in order to prevent the participants from having beforehand knowledge about the texts which have been used for evaluation.

The texts were manually tagged within the sentence frame. Both subsets of the corpus contain complex sentences, which may be considered as a bias in the experiment. The total amount of reference tagged text is about 140 000 words. Both genres are roughly equally represented.

The training corpus has been globally distributed to the participants in January 1996, while the dry-run corpus was distributed individually to the participants in fall 1996. The test corpus has been distributed individually at the end of December 1997.

6. Tagging the Reference Corpora

As in GRACE, the basis for evaluation are the PoS tags attributed by an expert in a reference corpus of literary and journalistic texts, an agreement must be reached, in particular as far as the definition of a reference tagset is concerned as well as the definition of adequate means to allow the effective tagging of the the text selected to serve as reference material for the evaluation.

The Reference Lexicon A reference lexicon was first thought necessary as a useful resource (1) for the participants with insufficient lexical data, and (2) for the human expert in charge of the tagging of reference corpora. It was indeed initially assumed that the reference tagging would be performed on the basis of tags extracted from the reference lexicon.

We therefore studied the availability of existing French lexica, such as: INTEXT (LADL) (Silberstein 1993), BREFLEX and BDLEX which have both been build in the framework of the “GDR-PRC Communication Homme-Machine”, lexica resulting from in-house efforts of the organizers (the electronic thesaurus extracted from the “Trésor de la Langue Française” (INaLF) and the lexicon of the École Nationale Supérieure des Télécommunications (ENST)), and lexica resulting from European Union funded projects like MULTEXT (Veronis and Ide 1994).

Notice that the goal was not to create new lexica from scratch but to see how existing ones can be re-use, extended, merged, etc.. The resulting lexicon (called the GRACE/MULTEXT lexicon) contains about 310 000 entries corresponding to 232 000 distinct word forms and 29 000 lemmas.

In fact, the GRACE/MULTEXT lexicon was not distributed to participants. The main reasons for this were: (1) the constant evolution of the reference tagset during the campaign and the low speed of the lexicon update process (the up-to-date version of the lexicon was always late and thus never really available); (2) most of the participants preferred to work with their own lexical resource (only one of them actually asked for the reference lexicon).

However, the non availability of a reference lexicon for the participants was finally considered as positive because it gave to the participants greater freedom for their strategy concerning segmentation and compound words identification.

In addition, the reference lexicon was effectively used internally in GRACE, as a reference for the tokenizing and the tagging of the reference corpora in the first two phases of the action (training and dry-run) (see “Tagging the corpora” hereafter).

Tagging the corpora As already stated two corpora were used: the FRANTEXT corpus texts (INaLF) contains mainly literary texts from the beginning of the century. The second corpus is constituted of contemporary extracts of the newspaper ”Le Monde”. The total amount of reference tagged text is about 140 000 occurrences (words). 30 000 of them are lemmatized (phases 1 and 2). 110 000 are not lemmatized (phase 3).

For the manual tagging, automatic tools were used to segment and/or pre-tag texts which were then manually revised in context. For the two first phases (training and dry-run), initial pre-processing was performed by means of a specific tool designed by the GRACE committee. This tool segmented the text according the GRACE conventions and assigned to each of the tokens the corresponding reference tags extracted from the reference lexicon. The text (presented with one token per line associated with a lemma and

a sequence of GRACE tags) was then manually revised for disambiguation and/or rectification.

The issue of the un-completeness of the lexicon arose. The lexicon provides lists of potential tags which do not necessarily fit for tagging in context, where function prevails. Because of the important number of transcategorization phenomena (“category shifts”), the preprocessing tool based on the reference lexicon appeared not to be adequate. For this reason, in the third phase of the project, reference texts were first tagged with a version of the Brill tagger trained on French texts (with a specific tagset developed by Josette Lecomte at INaLF), the result being mapped onto the GRACE tagset before being manually validated.

Only one person was in charge of the tagging task. No cross-tagging was possible, nor cross-revision for the dry-run, because of drastic cuts in the GRACE budget. Only a subset (about 10%) of the tagged corpus was submitted to cross-revision for the final test phase.

For the two first phases, no statistics were established concerning time spent in the various manual operations. For the third phase, the total human time spend to map, disambiguate, and revise, was estimated to about 360 person*hours (which corresponds to a processing speed of about 6 words per minute).

7. The current state of the program

During the first phase (training) a corpus of around 10 millions words of texts evenly distributed between literary excerpts and newspaper articles was distributed to the participants (January 96).

The second phase (dry-run) started with the tagging by the participants of a corpus of roughly 450,000 forms. During a workshop held in coordination with the JST97 (Journées Scientifiques et Techniques du Réseau Francil, April 1997, Avignon, France), the first intermediary results for 15 out of the 20 participating systems were presented to the participants in the presence of two external observers. The results were only presented for absolute and relative evaluation as they are the least expensive to compute and the calendar did not allow us to compute the results for clustered evaluation. 3 participants did not attend the workshop and 2 had already retired from GRACE, one formally, the other by not marking the dry-run corpus.

The initial participants to GRACE are: ATT Bell Labs. (USA), GREYC-URA 1526 (FR), INGENIA (FR), CRISTAL (FR), IAI (D), CNET (FR), XRCE (FR), LATL (CH), LIA-LPL (FR), TGID (FR), ISSCO (CH), SYNAPSE (FR), CLIPS (FR), ILR-IMS (D), IBM (FR), MEMODATA (FR), GSI-Erli (FR), CITI (CA), INaLF (FR) and LIMSI (FR).

The third phase (tests) was a re-run of the dry-run with the same protocol but new data. Among the GRACE participants only 13 have returned the test corpus tagged and are still considered to be active participants. The results of the systems presented by the organizers which will necessarily be public because of their involvement, and will not be officially considered in the final results of the evaluation campaign for obvious reasons of fairness. Among the

13 remaining participants, 4 took the option to participate anonymously, while all the other accepted to have their results identified.

The participants originate from both research and industry and are from various nationalities (France, Germany, Switzerland, Canada, and United States). The systems presented also strongly vary in their nature, ranging from statistical systems to rule based system, even including full-fledged syntactic analyzers with functionalities specifically dedicated for Part-Of-Speech tagging.

Right now, the test data returned by the participants have been re-aligned and the computation of the results (only absolute performance evaluation for now) has begun and should be ready for the coming workshop (scheduled to happen mid May 98) which will mark the beginning of the adjudication phase for the tests. A point of unknown still subsists about the protocol, because up to know, we did not had the resource for computing the results for the clustered evaluation scheme applied to real size data, having to restrict to demo runs on limited data samples. Hopefully this point will be resolved during the test adjudication.

In the line of the work done for English in the AMAL-GAM project at Leeds University, the valorization the by-products of GRACE done by building a tagged corpus of approximately 1 million words from the data tagged by the participants during the dry-run and the tests has already started and received support under the name of the MULTITAG project in the context of the joint program "Ingénierie des Langues" (Language Engineering) between the SPI (Engineering Sciences) and SHS (Human Sciences) departments of the CNRS.

8. Related Programs

In one of its programs, the Aupelf-Uref has set up a network for French language engineering (FRANCIL), co-ordinated by Joseph Mariani. One of the aims of this network is the creation and distribution of language resources for French, and the evaluation of natural language processing systems and methods used in language engineering. A large number of laboratories from French speaking countries are contributing to this network. In 1994, the Aupel-Uref has published a series of calls for tenders concerning concerted research programs (ARC-Actions de Recherche Concertées) centered on the Evaluation Paradigm. Two main directions have been identified :

- line A: linguistics, computer science and written corpora which addresses issues related to the development of systems for message routing (A1), bi- and multi-lingual corpus alignment (A2) (Veronis 1996), automatic terminology extraction from corpora (A3), and text understanding (A4),
- line B: linguistics, computer science and oral corpora, with subtopics like dictation (B1) (Bimbot 1996), spoken dialog (B2), and speech synthesis (B3).

GRACE collaborates with the SILFIDE (Bonhomme 1996) national project (Serveur Interactif pour la Langue

Française, son Identité, sa Diffusion et son Étude) started in 1996 and co-funded by Aupelf-Uref and CNRS. The goal of this project is to organize a network of data servers for language resources for the study of French. SILFIDE does not aim at the integration of existing resources (corpora, lexica and tools) but intends to provide informations on what is available and under which conditions, in a standardized format using French as support language. SILFIDE will provide WEB support for distributing the byproducts of the GRACE evaluation which will be finalized in MULTITAG (see end of previous section). For a CDROM distribution we intend to rely on ELRA (Choukri 1996) (the European Language Resources Association).

At the European level, the recently begun preparatory action ELSE which proposes to address the topic of evaluation by aiming at the preparation of a general infrastructure for Language Engineering evaluation in the context of the fifth Framework Program, will re-use a lot of the experience and tools developed in the course of GRACE. The general infrastructure ELSE intend to develop concerns the definition of a (semi)-automatic and task-independent protocol framework for black-box and quantitative evaluation of Natural Language Processing (NLP) systems in a multilingual set-up build around the concept of "control task".

9. Conclusion

For the first time, GRACE succeeded in using the task of Part Of Speech tagging to compare the approach of 13 different systems using a common formalism and real-life data in the course of an open evaluation campaign.

The evaluation by-product, a corpus of 1 million words tagged by 13 different systems (out of which 450,000 forms have been tagged by 18 systems) with tags mappable into a common formalism is in itself also a valuable result of the GRACE campaign. The ongoing comparative evaluation of this basic data, as well as the potential system improvement due to competitive stimulation and exchange of ideas, constitute an additional substantial contribution to the field of tagging and tagger evaluation (following up on GRACE, one of the industrial participant already added a new product to his catalogue).

In addition to these concrete results, the whole evaluation protocol (including lexicon and corpus collection, formatting and testing, tagset definition, project organization, and data exchange procedures) represents a precious know-how that makes GRACE the first successful large-scale experiment aiming at the definition of a quantitative black-box evaluation procedure for taggers, which resulted from a close collaboration between computer scientists and linguists.

More generally, GRACE represents a good illustration of the kind of efforts necessary for the Language Engineering community to progress towards the definition of an operational evaluation methodology for Natural Language Processing systems.

10. References

- Adda G., Blache Ph., Mariani J., Paroubek P., and Rajman M.. Action GRACE - Mise en place du paradigme d'Évaluation - Application au domaine de l'analyse morpho-syntaxique. In *Proceedings of the Conférence sur le Traitement Automatique du Langage Naturel (TALN'95)*, Marseille, France, juin, 1995.
- Bimbot F. 1996. Un point sur les Actions de Recherche Concertées (ARC) - Thème B1. In *Lettre d'information de l'Aupelf-Uref*, N. 3, Avril.
- E. Black et al., "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars", *ARPA HLT Wokshop*, Mars 1991.
- E. Black, "Parsing English by Computer: the State-of-the-Art", *International Symposium on Spoken Dialog*, Tokyo, November 1993.
- E. Black, "A New Approach to Evaluating Broad-Coverage Parsers/Grammars of English", *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP'94)*, UMIST, Manchester, September 1994.
- Patrice Bonhomme, Florence Bruneseaux, Jean-Marie Pierel, Laurent Romary, "Vers une normalisation des ressources linguistiques: le serveur SILFIDE", Actes des 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'Aupelf-Uref, Avignon, Avril 1997.
- M. Cavazza, "Méthodes d'évaluation des logiciels incorporant des technologies d'informatique linguistique", *rapport MRE-DIST*, 1994.
- Khalid Choukri, "Activités d'ELRA, Association Européenne pour les Ressources Linguistiques", Actes des 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'Aupelf-Uref, Avignon, Avril 1997.
- Richard Crouch, Robert Gaizauskas, Klaus Netter, "Interim Report of the Study Group on Assessment and Evaluation", EAGLES, Draft report, march 1995.
- EAGLES, "Evaluation of Natural Language Processing Systems", *EAGLES report EAG-EWG/PR2*, July 1994.
- Harman Donna K. (ed.), "The First Text REtrieval Conference (TREC-1)", *NIST Special publication 500-207*, National Institute of Standards and Technology, Gaithersburg MD., 1993.
- Harman Donna K. (ed.), "The Second Text REtrieval Conference (TREC-2)", *NIST Special publication 500-215*, National Institute of Standards and Technology, Gaithersburg MD., 1994.
- R. Hauser, "Results of the 1. Morpholympics", *LDV-FORUM*, vol. 11-1, June 1994, ISSN 0172-9926.
- R. Hauser, "The Coordinators' Final Report on the First Morpholympics", *LDV-FORUM*, vol. 11-1, June 1994, ISSN 0172-9926.
- G. Leech, A. Wilson "EAGLES Morphosyntactic Annotation, Draft - Work in Progress". *Draft technical report*, Lancaster, EAG-CSG/IR-T3.1., October 1994.
- MUC-4, *Proceedings of the Fourth Message Understanding Conference*, Morgan Kaufman, 1992.
- Patrick Paroubek, Gilles Adda, Joseph Mariani, Martin Rajman, "Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de Parties du Discours pour le français", Actes des 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'Aupelf-Uref, Avignon, April 1997.
- M. Silberztein, *Dictionnaires électroniques et analyse automatique de textes - Le système INTEX*, Masson, Paris, 1993.
- K. Sparck Jones, J.R. Galliers, *Evaluating Natural Language Processing systems*, Springer, 1996.
- Beth Sundheim, "Third Message Understanding Evaluation and Conference (MUC-3): Phase 1 status report", *Proceedings of the Speech and Natural Language Workshop*, Morgan Kaufman, Pacific Grove, CA, February 1991.
- Simone Teufel, "A Support Tool for Tagset Mapping", *Proceedings of the Workshop SIGDAT (EACL95)*, 1995.
- Nancy Ide, Jean Veronis, "MULTEXT: Multilingual Text Tools and Corpora, Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan, 1994.
- Jean Veronis et al., "Common Specifications and Notation for Lexicon Encoding", *rapport MULTTEXT LRE-62-050*, WP1.6 Deliverable, version préliminaire 1994.
- Jean Veronis, "Un point sur les Actions de Recherche Concertées (ARC) - Thème A2", Lettre d'information de l'Auplef-Uref, N. 4, July 1996.