

# Porting to an Italian Surface Realizer: A Case Study

Alessandra Novello and Charles B. Callaway

ITC-irst, TCC Division

via Sommarive, 18

Povo (Trento) I-38050, Italy

{callaway, novello}@itc.it

## Abstract

Multilingual generation is becoming an increasingly important aspect of implemented systems that showcase the abilities of generation systems. Most such systems require multiple grammars, one for each language which must be deployed. Yet little is known about the development costs for additional languages which are developed not from scratch, but by adapting existing resources. We ported a standard English surface realizer and grammar with wide coverage to Italian. After describing major grammatical differences, we quantitatively specify the porting process and present statistical information for the changes we found necessary to develop the new grammar.

## 1 Introduction

Multilingual generation systems will play increasingly important roles in showcasing the abilities of deep NLG (Paris et al., 1995; Stede, 1996; Callaway et al., 1999; Scott, 1999). These systems require an array of resources that can function regardless of the language selected, such as discourse and sentence planning rules, lexica, and pronominalization strategies. One of the most important of these resources is the grammar that a surface realizer uses to produce linearized text from a syntactic sentence plan, and multilingual

systems must use a distinct grammar for each desired language.

While many multilingual systems have either developed grammars from scratch or borrowed them from other projects, relatively few projects have focused on reworking existing grammars to port them to new languages. Most such work has been connected with the KPML environment (Bateman, 1997; Aguado et al., 1998; Kruijff et al., 2000), and the newer EXPRIMO system developed at Edinburgh and based on ILEX (Oberlander et al., 1998). However, these projects have not addressed the issue of exactly how much effort is involved in converting a surface realizer for one language into another in a quantitative manner. And while (Callaway et al., 1999) presented basic data on an English to Spanish project, it was not comprehensive enough to allow future projects to accurately estimate what potential development costs might be.

A separate trend has been to justify as both useful and cost-effective the continued use of resources in investigating deep natural language generation over other, more near-term approaches such as template generation. In order to make an informed comparison, hard data is needed on the costs for developing and maintaining projects which use both formalisms. In this article, we provide such data for the grammar and morphology development of an Italian surface realizer as a first step in allowing such comparisons to be made.

During the course of work on a multilingual generation system for English and Italian, we took elements from both the original FUF/SURGE

systemic-functional surface realizer for English (Elhadad, 1991; Elhadad, 1992; Robin, 1994) as well as a less-developed Spanish version (Callaway et al., 1999) of that same realizer to create a new Italian version<sup>1</sup>. The porting process involved changes to morphology, linearization, and the grammar, while leaving unchanged other features of the FUF system such as formatting and efficiency directives. This paper presents the results of creating the new surface realizer, including an overview of differences between the languages and a quantitative analysis of the effort and changes involved.

## 2 Examples of Language Differences

The differences between Italian and English are not significant compared to languages from differing families. The following areas are indicative of the types of linguistic changes necessary when generating Italian text as opposed to English. Extensive catalogues of such changes for other languages such as French also exist (Rayner et al., 1996). The various categories for Italian include:

**Morphology** Changes that affect the prefixes and suffixes of words for purposes of agreement, along with interactions between surface forms after they have already been syntactically specified.

- *Irregular Words*: Irregulars mainly concern lexical forms for nouns, verbs, and adjectives (which have few irregular forms in either English or Italian). Besides the three most important, regular rules for Italian noun pluralization (-o/-i, -a/-e, -e/-i), there are more than 20 other minor rules for pluralization (e.g., nouns with accented endings: sing. *crisi*, pl. *crisi*) and a third category of completely irregular plurals (e.g., sing. *tempio*, pl. *templi*). Furthermore, while English can express all verbs with at most five basic forms plus auxiliaries, Italian verbs can have up to 49 different irregular forms.
- *Contractions*: Italian can form contractions between a preposition and a definite article,

such as *su + la* ⇒ *sulla* (“on” + “the”). Additionally, contractions can occur between certain proclitic pronouns and verbs beginning with a vowel or ‘h’ plus a vowel (e.g., *l’ho vista* “I have seen her”, or *c’è* “there is”). There are also rules for dropping unstressed vowels, especially after infinitives: *aver detto* rather than *avere detto*, or with enclitic pronouns: *fare+lo = farlo* “to do it”.

**Word Order** Differences in relative positioning of certain syntactic categories with respect to others and co-occurrence constraints.

- *Adjectives*: Adjectives in Italian can be found in pre-nominal or post-nominal position. Some adjectives allow only one position, so a feature “pre-n” or “post-n” must be added to the lexicon. Lots of adjectives can appear in both positions, causing the distinction between their appositive and restrictive use. Since some adjectives change their meaning completely according to their position,

|                          |                          |
|--------------------------|--------------------------|
| <i>La vecchia strada</i> | <i>La strada vecchia</i> |
| (lit. the old street)    | (lit. the street old)    |
| “The familiar street”    | “The old street”         |

this requires that they be listed as different lexical items. Further order constraints arise when more than one adjective determines a nominal head:

|                                  |
|----------------------------------|
| <i>Un nuovo cinema italiano</i>  |
| (lit. a new cinema Italian)      |
| <i>Un cinema italiano nuovo</i>  |
| (lit. a cinema Italian new)      |
| <i>*un cinema nuovo italiano</i> |
| (lit. a cinema new Italian)      |
| “A new Italian cinema”           |

- *Subject Position*: Subject in Italian can occur either in preverbal or postverbal position. It generally precedes the verb, but it follows it with unaccusative and unergative structures:

*E’ arrivata Laura.*  
(lit. Is arrived(agr.) Laura.)  
“Laura arrived.”

Other verbs such as “mancare” (be lacking), “piacere” (be pleasing), and “servire” (be of

<sup>1</sup>The resulting grammar is freely available for research purposes at <http://tcc.itc.it/>

use) strongly prefer the inversion of complement order:

*A Luca piace la pasta.*  
(lit. (dat-prep.) Luca likes the pasta.)  
“Luca likes pasta.”

The postposition of the subject is also required with interrogative WH:

Che cosa ha comprato Giorgio?  
(lit. What thing has bought Giorgio?)  
“What did Giorgio buy?”

- *Clitics*: Accusative clitics precede the finite verb, while direct objects usually follow it:

*Mary l’ha letto.*  
(lit. Mary cl.(acc) has read.)  
“Mary has read it.”

With restructuring verbs, clitics can attach either to the main verb or to the infinitive:

*Mary vuole comprarlo.*  
(lit. Mary wants to buy cl.(acc).)  
*Mary lo vuole comprare.*  
(lit. Mary cl.(acc) wants to buy.)  
“Mary wants to buy it.”

But clitics follow the verb when the mood is imperative:

*Lo regali a Gianni.* (indicative)  
(lit. cl.(acc) give to Gianni.)  
“Give it to Gianni.”

*Regalalo a Gianni!* (imperative)  
(lit. give cl.(acc) to Gianni!)  
“Give it to Gianni!”

When both dative and accusative clitic are required, the order of complements is inverted (dative precedes accusative):

*Mary me lo dice.*  
(lit. Mary cl.(dat) cl.(acc).)  
“Mary tell it to me.”

**Grammar** Modifications to choosing which syntactic categories are allowed in which positions and what defaults are given to individual features.

- *Secondary Clauses*: Sentences where matrix verbs govern a gerund clause, such as

*Ho visto il ragazzo uscendo dalla chiesa.*  
(lit. have seen the boy leaving from the

church.)  
“I saw the boy leaving the church.”

by default prefer to keep the subjects identical, whereas in English the object of the matrix verb generally corefers to the subject of the matrix verb by default. Thus where the boy was leaving the church in the English example above, in the Italian version it is the speaker who was leaving the church.

- *Formal/polite pronouns*: Italian uses the third person feminine address “Lei” (even when it is addressed to a male person) instead of the second person. The use of the polite form involves changes to verbs and pronouns when the mood is imperative. Indeed, Italian has imperative forms for the second singular person and second and third plural, but changes to the subjunctive for polite imperatives, eg.:

*Leggi!* (imperative)  
“Read!”

*Legga!* (imperative realized by a subjunctive)  
“Read!”

Further changes arise from the use of clitics:

*Leggilo!* (enclitic in familiar form)  
“Read it!”  
*Lo legga!* (proclitic in polite form)  
“Read it!”

- *Verb-governed pronouns*: Most notably, dative constructions in Italian are much different than those in English. Features in SURGE like “dative-shift” are not useful and are thus not referenced in the Italian Grammar.

**Discourse** Differences in which although a constituent is allowed grammatically, one language prefers something slightly different.

- *Zero pronominalization*: Also called *pro-drop* (Haegeman, 1994), this is the result of not mentioning a repetitive subject pronoun, as it is redundant given that verbs are inflected for a subject’s number and gender (Di Eugenio, 1998).

### 3 Coverage of the Italian Grammar

Most symbolic generation systems use regression testing as a means of demonstrating the amount

|  |   |
|--|---|
| <pre>"This car is expensive." ((cat clause)  (proc ((type ascriptive)         (mode attributive)))  (partic ((carrier ((lex "car")                     (cat common))                 (distance near)))           (attribute ((lex "expensive")                     (cat ap))))))</pre> | <pre>"Questa macchina e' costosa." ((cat clause)  (proc ((type ascriptive)         (mode attributive)))  (partic ((carrier ((lex "macchina")                     (cat common)                     (gender feminine)                     (distance near)))           (attribute ((lex "costoso")                     (cat ap))))))</pre> |
|--|---|

Figure 1: A simple example with almost direct feature-feature mapping

of coverage of a particular language. For example, the FUF/SURGE surface realizer includes over 500 examples of paired inputs and outputs covering a wide range of phenomena subdivided into categories such as yes/no questions, relative clauses, noun phrases, etc.

Although we did not attempt to duplicate coverage for this extensive test suite, we did obtain enough coverage to allow for the production of multiple paragraphs of simple text.

Throughout our efforts, we strove to make the input representation as similar as possible to the existing SURGE test suite. An example of this similarity is found in Figure 1, where only individual lexical items differ between the two functional descriptions. Thus the new surface realizer can be used with existing discourse and sentence planners, with only changes to the lexicon needed in a pipelined NLG architecture.

Figure 2 shows a more complex example where the structures of the sentences are so divergent that either the sentence planner must be able to generate different sentential representations or the interface to the surface realizer must be moved even higher to exclude all syntactic references. For Italian and English we have seen a higher proportion of the simple cases in our application environment, although it is highly likely that the exact proportion changes by language pair.

#### 4 The Porting Process

Porting the SURGE grammar to Italian was accomplished in a systematic way. We first worked on the morphology of individual words in isolation.

Italian words typically have much more inflection than those in English as is documented in most books on language instruction. Italian morphology is well defined, and thus we used such materials to ensure that morphologic coverage was complete and could be performed rapidly. Another quick change that could be made was to replace all lexicalized closed-class words in the grammar (such as the English “to” with infinitives).

Next, we performed basic testing of examples in the provided SURGE test suite (with Italian lexicalizations substituted and additional features like gender added) to gauge how many changes might be necessary. The results showed that morphology interactions and linear precedence were the most obvious errors that were immediately noticeable. We thus proceeded to attempt to fix these errors before concentrating on the grammar itself. After reexamining the newly regenerated sentences, we found that most were recognizably similar to the Italian equivalent “gold standards”, even if they contained many errors.

We next controlled for morphological interactions between adjacent words (similar to contractions in English), adjust for accented characters not present in English, and add simple feature propagations such as gender in predicate verb and attributive noun constructions where similar such features (*e.g.* number) already existed. Finally, we began work on the more difficult differences in actual grammar, which required significantly more time than the changes mentioned above.

At this point there are two possible directions that a surface realization project can take: to continually develop the grammar as a linguistic-

```

"The town is meant to be Trento"
((cat clause)
 (proc ((type lexical)
        (lex "mean")
        (voice passive)
        (subcat ((1 {^3 lex-roles influencer})
                 (2 {^3 lex-roles influenced})
                 (3 {^3 lex-roles soa})
                 (1 ((cat np)))
                 (2 ((cat np))
                    (3 ((cat clause)
                        (mood to-infinitive)
                        (controlled {^ oblique 1}))))))))))
 (lex-roles ((influenced ((cat common) (lex "town"))
                 (soa ((proc ((type equative))
                            (partic ((identified ((cat proper)
                                                    (lex "Trento"))))))))))))

"Si ritiene che la citta' sia Trento"
((cat clause)
 (proc ((type lexical)
        (lex "ritenere")
        (subcat ((1 {^3 lex-roles believer})
                 (2 {^3 lex-roles belief})
                 (1 ((cat np)))
                 (2 ((cat clause)
                    (binder ((lex "che")))
                    (mood bound))))))
 (lex-roles ((believer ((cat personal-pronoun) (case reflexive)
                        (animate yes) (person third))
                 (belief ((cat clause)
                          (proc ((type ascriptive) (mode equative)
                                   (mood subjunctive)))
                          (partic ((identified ((lex "citta") (cat common)
                                                (gender feminine)))
                                      (identifier ((lex "Trento")
                                                  (cat proper))))))))))

```

Figure 2: A more complex example where features are not aligned

only initiative to provide extensive coverage (the breadth approach), or to begin to flesh out particular projects and provide only the type of linguistic structures they need but in greater detail (the depth approach).

## 5 Quantitative Results

The overall process required approximately five person-months, split between two people: one with pre-existing knowledge of FUF/SURGE and a native English speaker, the other without knowledge of FUF/SURGE who is a native speaker of Italian. By the middle of the project, the second person was capable of making major grammatical

changes unaided. Also, as the project continued and intensive knowledge of Italian was increasingly necessary, the burden of the labor shifted to the native Italian speaker. Below we detail the information gathered after this five month period, when the surface realizer was sufficiently developed to produce a paragraph of Italian in a working demo where the equivalent English paragraph was also generated from the same discourse plan.

Table 1 shows various quantitative aspects of the grammar-creation process. *Lines* refers to the number of lines of actual code devoted to different items. While not indicative of the amount or degree of changes necessary, the results show a high degree of correlation between the overall sizes.

|                    | English Grammar |             | Italian Grammar |             |           |
|--------------------|-----------------|-------------|-----------------|-------------|-----------|
|                    | Lines           | Constraints | Lines           | Constraints | Work Time |
| Top Level          | 477             | 392         | 496             | 335         | 1%        |
| Adverbials         | 1191            | 3167        | 1034            | 2861        | 2%        |
| Clauses            | 461             | 202         | 432             | 200         | 8%        |
| Conjunctions       | 293             | 126         | 283             | 100         | 3%        |
| Determiners        | 900             | 688         | 933             | 636         | 17%       |
| Clause Mood        | 511             | 282         | 485             | 245         | 2%        |
| Noun Phrases       | 753             | 479         | 732             | 463         | 17%       |
| Transitivity       | 749             | 420         | 743             | 423         | 5%        |
| Verb Group         | 927             | 723         | 699             | 453         | 6%        |
| Clause Voice       | 432             | 235         | 519             | 278         | 8%        |
| Grammar Totals     | 6694            | 6714        | 6356            | 5994        | 69%       |
|                    | Lines           |             | Lines           | Changed     | Work Time |
| Irregular Verbs    | 210             |             | 450             | 100%        | 10%       |
| Other Irregulars   | 40              |             | 80              | 100%        | 7%        |
| Other Morphology   | 680             |             | 700             | 30%         | 5%        |
| Linearization      | 640             |             | 640             | 5%          | 4%        |
|                    | Lines           | Examples    | Lines           | Examples    | Work Time |
| Regression Testing | 5922            | ~500        | 680             | 45          | 5%        |
| Totals             | 14186           |             | 8906            |             | 100%      |

Table 1: Grammar, Code, and Resources Data

This indicates that even when substantial changes were made, they effect was to replace rather than increase or decrease the size of the grammar.

A more closely related statistic is the number of actual *constraints* incorporated in the grammar. Due to the feature-based nature of the functional unification formalism (Kay, 1979) underlying SURGE, it is possible to count the number of times features are expressed. This ignores the effects that comments and formatting imposed by different grammar authors have on the grammar itself as a data structure.

*Work time* reflects the percentage of the five person months that were spent in certain areas of the grammar and other resources. Small percentages indicate that a grammar module was little changed from the previous grammar. Unlike the other statistics, this is an estimate, as we did not count the actual time spent in each area. Importantly, it is probably not possible to make a completely accurate estimate of time spent, as different people work at different speeds, and even a sin-

gle person will work slowly or quickly on different days.

This data indicates that noun/determiner phrases and morphology required the most amount of work. We hypothesize that this data also indicates the verbal systems between English and Italian are closer than the nominal systems.<sup>2</sup>

To ensure the robustness and to double check the coverage, we employed the grammar as part of several multilingual projects we are currently researching. Thus the size of the regression test set is rather small compared to that of the English version, especially as we did not find examples of questions, appositions, partitives, dates, forms of address, *etc.* that are contained in the original regression set. From this work we estimate that another five to six person months would be necessary to ensure that these additional types of possi-

<sup>2</sup>The large differences in the *verb-group* module are not indicative of significant change; we removed many of the extensive tenses SURGE has for sentences like “He was about to be going to the bus.” With these included, the file length and number of constraints are still tightly correlated.

ble grammar inputs could also be generated.

## 6 Discussion

There are at least three important questions that need to be resolved in future research of this type:

- What impact is there on portability
- What type of regression testing is needed
- How does porting for deep generation compare with that for template generation

It is highly desirable that multilingual generation can take place with a minimal amount of changes to architectural modules. Because the functional unification formalism is feature-based, features are a necessary aspect of the representation. But it is an open research topic whether different languages can be described with a similar set of semantic features (*c.f.* research on interlinguas) and if so, can such a representation generate a large enough set of paraphrases in each target language. Another way of looking at this is to ask whether the divergent structures in Figure 2 can be resolved so that they are identical but generate the different required syntax. If not, this problem must be pushed further up the NLG pipeline. In the texts we have generated, we have yet to find an example where changes above the level of the sentence planner must be made. In practice, most syntactic features which are not part of both languages (such as the dative-shift feature in SURGE) have been safely ignored because they are not referenced or constrained in the Italian grammar.

A second aspect is the use of regression testing with a small number of examples to “test” the coverage of the grammar. Without standards, regression testing is useless as a comparison metric between surface realizers because (1) it is not clear how many examples are necessary, (2) there is no recognized set of levels of coverage other than “not complete”, and (3) it is not clear how complex a set of examples needs to be (*e.g.*, does every possible combination of intermixed features or syntactic constructions need to be attempted).

Finally, it is important to be able to compare with a cost/benefit analysis the claimed advantages of linguistic-based (deep) generation over those of

string-based (template) generation. Related to this particular project are 4 aspects of this problem:

- What other NLG infrastructure was there to begin with? The Italian grammar we developed was used in a separate project which already had extensive NLG infrastructure. For both template and deep generation systems, most of the development effort goes into producing a working system for a single language. But it is unknown what the costs of additional languages are for each approach.
- How domain-independent is each approach? Intuitively, template approaches are highly domain-specific while deep generation is more domain independent. But to what extent has never been quantitatively demonstrated.
- How much effort is required to integrate with other NLG elements? For example, future work may find that pronominalization or revision algorithms have different effects depending on language. If hidden interaction complexities arise, they may have a large influence on cost/benefit ratios.
- What is the break-even point where development and maintenance costs for template approaches outweigh those for deep generation? In this project, would another 5 months of effort on the grammar result in a domain-independent surface realizer? We believe so, but detailed evidence must be collected on a template realizer and deep generation realizer working with an identical NLG pipelined system on an identical project and domain to be certain.

## 7 Conclusion

Multilingual generation is an increasingly important tool to demonstrate the widely-believed but little-substantiated intuition that natural language generation can provide effective and efficient systems whose development costs outweigh those of other methodologies such as template generation. But there has been little published evidence on what these costs may be, without which it is impossible to make an educated comparison.

We have thus presented a quantitative analysis of the effort required to build a grammar by reusing existing resources, a summary of the changes required, and estimates of how much work was devoted to varying aspects. This type of data is a necessary precursor to making future comparisons between differing methodologies on the basis of system development cost rather than traditional approaches which evaluate the text produced in a working system.

Finally, the material result of this project has been a functioning Italian generation grammar, which we plan to make available to the NLG community as an open source, freely available common resource.

## 8 Acknowledgements

This work was funded by the PEACH and TICCA projects, funded by the Autonomous Province of Trento. We would like to thank the anonymous reviewers, who gave us particularly lengthy and insightful comments.

## References

- G. Aguado, A. Bañón, J. Bateman, S. Bernardos, M. Fernández, A. Gómez-Pérez, E. Nieto, A. Olalla, R. Plaza, and A. Sánchez. 1998. ONTOGENERATION: Reusing domain and linguistic ontologies for spanish text generation. In *ECAI Workshop on Problem-Solving Methods and Ontologies*, Brighton, UK.
- John A. Bateman. 1997. Enabling technology for multilingual natural language generation: The KPML development environment. *Journal of Natural Language Engineering*, 3(1):15–55.
- C. Callaway, B. Daniel, and J. Lester. 1999. Multilingual natural language generation for 3D learning environments. In *Proceedings of the 1999 Argentine Symposium on Artificial Intelligence*, pages 177–190, Buenos Aires, Argentina.
- Barbara Di Eugenio. 1998. Centering in Italian. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering in Discourse*. Oxford University Press, Cambridge, MA.
- Michael Elhadad. 1991. FUF: The universal unifier user manual version 5.0. Technical Report CUCS-038-91, Department of Computer Science, Columbia University.
- Michael Elhadad. 1992. *Using Argumentation to Control Lexical Choice: A Functional Unification Implementation*. Ph.D. thesis, Columbia University.
- L. Haegeman. 1994. *Introduction to Government & Binding Theory*. Blackwell Publishers Ltd., Oxford, UK.
- M. Kay. 1979. Functional grammar. In *Proceedings of the Berkeley Linguistic Society*.
- Geert-Jan Kruijff, Elke Teich, John Bateman, Ivana Kruijff-Korbayová, Hana Skoumalová, Serge Sharoff, Lena Sokolova, Tony Hartley, Kamenka Staykova, and Jiří Hana. 2000. Multilinguality in a text generation system for 3 Slavic languages. In *COLING-2000: Proceedings of the 18th International Conference on Computational Linguistics*, Saarbruecken, Germany.
- J. Oberlander, M. O'Donnell, C. Mellish, and A. Knott. 1998. Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *The New Review of Hypermedia and Multimedia*, 4.
- Cécile L. Paris, Keith Vander Linden, Markus Fischer, Anthony Hartley, Lyn Pemberton, Richard Power, and Donia Scott. 1995. A support tool for writing multilingual instructions. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1398–1404, Montréal, Canada.
- M. Rayner, D. Carter, and Pierrette Bouillon. 1996. Adapting the core language engine to French and Spanish. In *Proceedings of NLP-IA-96*, Moncton, New Brunswick.
- Jacques Robin. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background*. Ph.D. thesis, Columbia University, December.
- Donia R. Scott. 1999. The multilingual generation game: Authoring fluent texts in unfamiliar languages. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, Stockholm, Sweden.
- Manfred Stede. 1996. *Lexical Semantics and Knowledge Representation in Multilingual Sentence Generation*. Ph.D. thesis, University of Toronto, Toronto, Ontario.