Computer Science Department, University of Corunna,
Campus de Elviña s/n, 15071 La Coruña, Spain
{vilares,grana}@dc.fi.udc.es

Spanish Philology Department, University of Santiago de Compostela,
Burgo de las Naciones s/n, 15705 Santiago de Compostela, Spain
fepili@usc.es

(*Received 25 January1997*)

Abstract

This paper describes an environment for the generation of non-deterministic taggers, currently used for the development of a Spanish lexicon. In relation to previous approaches, our system includes the use of verification tools in order to assure the robustness of the generated taggers. A wide variety of user defined criteria can be applied for checking the exact properties of the system.

Introduction. The choice of the FA model as operational formalism for tagging assures computational efficiency. Safety is guaranteed by the separation which exists between this operational kernel and the high-level descriptive formalism.

However, one of the major services of every lexicon ought to be to provide as much information as possible about errors, because of the complexity of actual implementations, and the natural evolution suffered by these kinds of systems. The goal is to minimize the time dedicated to debugging the system. For this reason, our discussion has a practical sense.

The approach presented allows verification of morphological analyzers by computing reductions. We are interested in verification methods combining modularity in the description of the system and flexibility in the verified properties, such as AUTO (Madelaine *et al.* 1989). AUTO computes small-scale models of finite transition systems, as is the case for FA's. These reduced systems are quotiens of that under study. The parameter of the reduction is a user-defined abstraction criterion, which embodies a particular point of view. One is therefore able to build a variety of quotiens of a same system, which are small enough to verify particular properties.

Spanish as a guideline example. We consider the case of Spanish as a running example. Spanish is an inflectional language, which shows a great variety of morphological processes, particularly non-concatenative ones. At present, we are able

to tag the most common 7000 lemmas of Spanish. The corresponding automaton has more than 35000 states and the average speed is 1400 words tagged per second. This complexity suggests the need to interface the morphological analysis with a formal proof system which allows us to verify easily the properties demanded.

Verification by reduction. The verification method we want to advocate in this paper is based on reductions of a global FA. These collapse states of the automaton to reach sizes reasonable enough to be outprinted and easily understood. We summarize some of the outstanding features we have implemented:

- **Tracing facilities.** From the global FA, the verification process lets the user obtain the path, that is, the set of states visited by the tagger, for a given word. This partial view allows us to check that the tagging is correct, and also to validate that the treatment units involved are working correctly.
- **Improving maintenance.** Latest versions should increase the power of previous systems, but the updating could unconsciously introduce erroneous patterns. Our goal is to detect these kinds of bugs in compile time. One way of doing this is to compare patterns. So, we can automatically take them out from the old tagger, which we assume to be correct, and verify whether they are present, or not, in the new model.
- **Automatic error recovery.** When an error makes the automaton stop the recognition process, the system automatically finds a *bridge transition* and catenates the portion of the erroneous string that we have identified as correct, with the endings corresponding to existing paths. As result, we obtain the set of possible corrections proposed by the verifier. It is clear that this strategy solves only the *change* error hypothesis. The error could of course be an *insertion* or a *deletion*, even in another position.

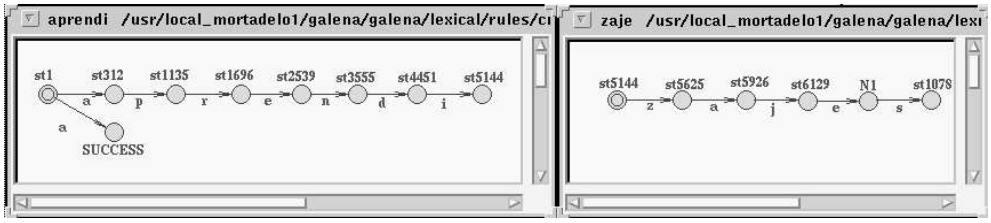

Fig. 1. Error recovery for `aprendizaje` (*learning*), from `aprendicaje` (an incorrect word).

Conclusion. The work described above is not closed. It represents only a first approach to verification in tagging, but preliminary results seem to be promising, and the operational formalism well adapted to deal with more complex problems.

References

Madelaine E. and Vergamini, D. 1989. : A verification tool for distributed systems using reduction of finite automata networks. In *Proc. FORTE'89 Conference*, Vancouver.