NATURAL LANGUAGE GRAMMARS FOR AN INFORMATION SYSTEM

Luis de Sopeña
CENTRO DE INFORMATICA
Universidad del Pais Vasco
BILBAO-SPAIN

ABSTRACT

The User Specialty Languages (USL) System is an applications independent natural language interface to a Relational Database System. It provides non DP-trained people with a tool to introduce, query, manipulate and analyse the data stored in a Relational Database via natural language. USL interfaces with different languages; in the present paper the grammar developed for Spanish is presented, and compared with the German grammar which was previously implemented and upon which it is based. Their main differences are pointed out, and the generality of the system to deal with other natural languages shown.

1. INTRODUCTION

The User Specialty Languages (USL) System is an applications independent natural language interface to a Relational Database. It was designed to provide non DP-trained users with a tool to introduce, query, manipulate and analyse the data stored in a Data Base. USL is placed on top of a Relational Data Base Management System, System R (1), and its function is to translate natural language input sentences typed in by the users into queries written in the formal query language of System R.

A main objective of USL was to achieve applications independence, i.e. to describe that subset of natural language (morphology, structural words, syntactic structures and their interpretation) that can be used in the context of database interrogation, and that constitutes a core to be used in different domains. At the same time USL aimed at obtaining interfaces for diferent languages using the same technology: the first version of USL was developed for German by H. Lehmann N. Ott and M Zoeppritz (4,5,6) at the IBM Heidelberg Scientific Center. Subsequently and using it as a basis a Spanish version was also written (8), as well as an English one.

The components of USL are: a parser, a grammar with a structural vocabulary, a set of interpretation routines associated to the grammar rules, and the Data Base Management System, System R. the system has been designed in such a way that only the grammar and the structural vocabulary need to be changed in order to shift from one language to another; the rest of the items are common and can be shared by all language interfaces.

In addition to the system components a subject-dependent vocabulary has to be defined by the user for each application together with his/her database relations and views.

The present paper centers around the German and the Spanish Grammars developed for USL. Section 2 describes briefly the grammar structure and scope. Section 3 deals with a comparison of both grammars and points out their main differences.

The work reported here was developed during the author's stay at the Heidelberg Scientific Center of IBM Germany in 1981.

2. GRAMMAR OF THE USL SYSTEM

This section reports the main features of the grammar accepted by the USL parser. Detailed descriptions of the German and Spanish versions can be found in references (9) and (8), respectively.

© 1983 ACM 0-89791-107-5/83/006/0075 $00.75

2.1. GRAMMAR STRUCTURE

The USL grammar is written in a modified Backus-Naur Form. Here we shall briefly describe its elementsand the structures it uses.

The basic elements are:

- PRIMITIVES: Letters, digits and special symbols; they have to be declared in the grammar. Example:

  '0123456789' PRIMITIVE<DIGIT>;

- CONSTRUCTS: Grammatical categories defined for verbs (VERB), nouns (NOMEN), adjectives (ADJ), prepositions (PREP), noun phrases (NP), local (ABL) and temporal (ABT) complements, verb and one or more complements (Verb Complex, VC) verb and all its complements (Sentence Kernel, SK), etc. Example of a Construct declaration:

  <VERB> CONSTRUCT;

- FEATURES: They qualify Constructs and belong to one of three types: integer (values are integer numbers), logical (values are + and -, or 1 and 0), and case (values belong to a set of socalled casevalues, which must also be defined. Examples:

  <SG> FEATURE LOGICAL;

  singular feature for nouns, adjectives, etc. of singular number.

  <LAB> FEATURE INTEGER;

  qualifies the different prepositions (construct PREP).

  <TYP> FEATURE CASE;

  qualifies verbs according to the complements they require. Values of TYP have to be defined as, for example:

  <NI> CASEVALUE;

  for intransitive verbs (Nominative only).

  <NAD>CASEVALUE;

  for verbs with Nominative, Accusative and Dative complements.

The rules of the grammar make use of the elements that have previously been declared. They have a special symbol to separate left and right-hand sides: members on the right are input elements; after application of the rule they are substituted by the elements specified on the left.

There are four types of rules characterised by the symbol used to separate both sides; here we shall describe only the two most important types:

<= Fixed Token rules: they are used to define vocabulary words, by creating constructs from strings:

NOMEN:+MAS,+SG:FPE-NOMEN('PAIS')
<= 'PAIS';

The features for Masculine and Singular are assigned to the construct defined as NOMEN, which spans the string 'PAIS' ('country'). FPE-NOMEN refers to the name of the relation to be accessed for interpretation of the noun, here the same word 'pais' is used.

<- Grammar rules proper: Constructs and strings may appear on the right-hand side, one or several constructs on the left. For example, the rule attaching an adjective to a noun could be:

<NOMEN:1:FPE-ADJ(2,1)><-
<NOMEN:(MAS=MAS(2))!(FEM=FEM(2)),
(SG=SG(2))!(PL=PL(2))><ADJ>;

where for simplicity only the tests for gender and number agreement have been included on the right-hand side NOMEN construct. The resulting construct retains the features of the original NOMEN via the parameter 1, and the semantic routine FPE-ADJ is called for interpretation of adjectives modifying nouns.

2.2. GRAMMAR SCOPE

The USL grammar describes a subset of natural language to be used in database interrogation; therefore many structures not important in this context have not been included. However, both the German and the Spanish Grammars are quite comprehensive; they provide a similar coverage of the language, which includes:

- Wh-questions
- Yes/no questions
- Commands
- Statements
- Negation
- Adjectives
- Genitive Attributes
- Appositions
- Noun Complements
- Relative Clauses
- Quantifiers
- Comparatives
- Coordination
- Possessive Pronouns
- Locative Adverbials
- Temporal Adverbials
- Functions of sum, maximun, minimun, average, number
- Use of variables and functions

3. THE SPANISH GRAMMAR: A BRIEF PARALLEL WITH GERMAN

The firts USL grammar was written for

German ; the Spanish grammar is based on it, therefore there are many rules similarly formulated, sometimes only with slight variants. In this section the main differences between both grammars will be pointed out, as it can be found after comparing the descriptions given in references (9) and (8).

3.1. MORPHOLOGY

USL Grammars make no distinction between morphology and syntax; the same kind of rules serves for both purposes.

Rules defining morphology are of course very different in both grammars: noun and adjective affixes, inflection of nouns and adjectives are much more simple in the Spanish grammar, as only plurals need be described for nouns, and feminines and plurals for adjectives, and no declensions exist. But at the same time these constructs provide less information, as only gender and number are marked, and nothing helps determine the case (unless a preposition is later attached), as occurs in German.

German must also account for special cases, like nouns with adjective inflection, and nouns and names without article, which are not necessary in Spanish. As a result, several features used in German are not defined in Spanish.

3.2. SYNTAX

In many cases rules are very similarly formulated, only small changes can be detected by looking at the features checked on the righ-hand sides, the ones set on the resulting constructs, and the semantic routines invoked.

a)
Adjectives in Spanish can be placed either before or after a noun. The difference is mainly stylistic, but sometimes also the meaning changes depending on the adjective position (e.g. 'hombre pobre' vs. 'pobre hombre', caballo grande' vs. 'gran caballo'). For the purpose of USL, the most usual, noun-adjective sequence is to be expected, even though rules are also supplied for the reversed sequence. In German the adjective-noun sequence is only allowed.

b)
Even if nouns are morphologically unmarked for case in Spanish, the grammar makes use of case features. Any noun phrase not preceded by a preposition is in principle candidate for subject (Nominative) and direct object (Accusative). The preposition 'a' may introduce á Direct Object (for persons or any kind of 'personified' thing - but this has not been considered, and the Accusative feature is set on all these Noun Phrases); 'a' may also introduce an Indirect Object: e.g. 'amar A Juan', 'vender algo A alguien'. The preposition 'para' introduces Indirect Objects and the Dative feature is set on these Noun Phrases. As to the Genitive case, it does not exist as such in Spanish.

c)
Rules for nouns with determiners are very simple in Spanish, only gender and number checks are needed. German has to restrict the case features too, and distinguish between nominal inflection, adjective inflection, and the determiners 'wieviel', 'lauter/nur', and 'der/ein'. Many more rules much longer and complicated are therefore necessary in German, while only three rules are needed in Spanish.

d)
Spanish and German use the same three gender features for masculine, feminine, and neuter. However, neuter nouns de not properly exist in Spanish as they do in German: the neutral article 'lo' determines adjectives used as nouns, like 'lo bello', 'lo util', 'lo mejor' (the beautiful', 'the useful', 'the best'). They always have a collective and abstract meaning which is not expected to occur in database interrogation. Therefore, apart from a rule attaching the article 'lo' to a noun defined as neutral (which the vocabulary definition programs allow), no further use is made of the feature in the Spanish grammar.
Other possible uses of 'lo' preceding relative clauses of prepositional phrases, e.g. 'lo que paso', lo de siempre', have not been considered.

e)
In the rules for the formation of the SENT construct (the one spanning the whole input string), one is provided to transform a declarative sentence into a yes/no question. This transformation in the interpretation of a statement as an interrogation is done if a '?' is found at the end of the declarative sentence. In Spanish, the freedom in word order makes it possible to ask a question using a linear order just by a change in the intonation; when typing, a question mark is required to indicate this change. For example, the sentence 'Maria vive en Bilbao' is a statement ('Maria lives in Bilbao'), but 'Maria vive en Bilbao?' is an interrogation ('Does Maria live in Bilbao?').

f)
The interpretation of negation in front of a Verb Complex or Sentence Kernel is special in Spanish. If the quantifier 'ningun' ('no') is in the Verb Group, no negation should actually apply but the one already implied by the quantifier. In this way the sentences

'No vive ningun empleado en Bilbao?'

'Ningun empleado vive en Bilbao?'

('Does no employee live in Bilbao?')

will get the same interpretation, even if the first one has an additional negation in front of the sentence.

g)
The rules for Verb Complex (VC) and Sentence Kernel (SK) formation are quite similar. Only in Spanish when a Subject Noun Phrase is attached to a VC, a feature called SUBJ is checked in both constructs, and the Noun Phrase checked for presence of the quantifier 'todos' ('all').
The first check is tentative and tries to avoid syntactic ambiguities between Subject and Accusative Object arising in some sentences. For example the sentences 'Exporta España vino?' and 'Exporta vino España' (which both mean 'Does Spain export wine?') have two parses, 'vino' being subject in one parse and Direct Object in the other. This occurs because the only knowledge USL uses for disambiguation is the one contained in the structure of database relations, and no real knowledge representation device guides the system. The database contents are not duplicated in a dictionary, and therefore every unidentified string found in the input sentence is interpreted as a database value. In our example neither 'España' nor 'vino' are defined in the vocabulary, both of them are values stored in the database, and the system without semantic knowledge is unable to discard 'vino' as subject of 'exportar'.

The feature SUBJ tries to characterize nouns and names and the matching verbs of which they only can be subject: here 'exportar' and 'España'. However the feature is not yet completely used because this would imply, among other things, the definition of the proper names stored in the database, which was up to now avoided by USL . If real use is to be made of the feature it will be necessary to define it as integer, and assign its different values to different groups of verbs and subject candidates, or to define different categories of subjects (e.g. persons, animals, countries, etc.) and specify for each verb the classes of subjects it requires. This will allow to perform a semantic typing inside the syntax.

The second check for a Noun Phrase with 'todos' is important when the Noun Phrase is attached as Subject to the right of a VC. Due to the USL left to right interpretation for the scope of quantifiers, if the NP is the first complement attached to the verb the interpretation will be correct. Otherwise the order of complements should be reversed so that the NP is moved to the first position. This is done by application of a special semantic routine. For example, the sentences:

'Compran todos los empleados coche?'
('Do all employees buy a car?')

'Tienen todos los paises capital?'
('Do all countries have a capital?')

are correctly interpreted, but

'Compran coche todos los empleados?'

'Tienen capital todos los paises?'

which have exactly the same meaning as before, would be interpreted as if all employees bought the same car, all countries had the same capital, due to the order of the quantifiers in the input sentence. Therefore the order of complements must be reversed for a correct interpretation of the sentences meaning.

h)
When the main verb is 'ser' or 'estar' ('be') special checks are sometimes necessary. If an adjective is attached as complement to the verb a gender check is needed between adjective and the other

Nominative, e.g.:

'Son casadas las secretarias?'

'Que secretarias son casadas?'

where the adjective 'casada' agrees with 'secretaria'. On the other hand, number agreement is avoided for coordinated Nominatives:

'Quien es/quienes son el jefe y la secretaria de Juan?'

i)
No preceding Genitive Atributes exist in Spanish, there are only genitives following their head nouns: 'salario de Juan' vs. two possibilities in German: ' Gehalt von Meier', Meiers Gehalt'.

j)
When defining Prepositions the German grammar must duplicate many of them to account for their use with Accusative of Dative noun phrases. In Spanish this is not necessary as no case needs to be associated with prepositions, only one definition is needed.

Rules have to be provided in Spanish for any sequence of complements after a verb requiring prepositional complements, even if actually some of the structures thus obtained would sound odd and will probably not be used, but they are grammatically correct.

k)
To describe Local Adverbials less rules are needed in Spanish, as less particular cases need to be accounted for. They allow for the formation of interrogatives using 'donde' ('de donde', 'hasta donde'), and for adverbs with more than one preposition, like 'desde encima de', 'por debajo de'.

i)
For Temporal Adverbials there are many

differently formulated rules. For example, the following formats are allowed for a date in the Spanish grammar:

4/9/1956
4-IX-1956
04.09.1956
cuatro de septiembre de 1956
4 de septiembre de 1956
26 de enero

enero de 1.979
año 1979
año 1979
mes de enero
mes de enero de 1979
dia/viernes 26 de enero

A date can also be attached to a time expression, e.g.: 'el 5 de agosto a las 16:20'.
Care must be taken to avoid ambiguities between two possible uses of the word 'horas' as a measure of a time interval (meaning 'hours'), and as a point in time (meaning 'o'clock'):

'dentro de 3 horas'
'durante 2 horas y 10 minutos'
'antes de las 5 horas'
'a las 3 horas 10 minutos'

More than one preposition is allowed to introduce a temporal adverbial:
'desde antes del martes',
'hasta despues de las 8'.

m)
The syntax of Relative Clauses is quite different in Spanish, due to the particular properties and uses of Spanish Relative Pronouns. Also, as Relative Clauses are subordinate, word order is different in German and the grammar must be provided with a special section to describe these sentences with verb final word order. This is not necessary in Spanish because all clauses, either main of subordinate, have always a largely free word order.

Most of the Spanish relative pronouns convey little information about the characteristics of their referent noun: 'que' can have a referent of any gender and number, 'cual' and 'quien' are only marked for singular, 'cuales' and 'quienes' for plural, unless an article is placed in front of the relative:

'Paises a los que exporta Italia'
('Countries to which Italy exports')

'Empleados de los cuales Juan es jefe'
('Employees of whom John is manager')

The last relative pronoun 'cuyo' ('cuya ', ' cuyos', 'cuyas': 'whose') admits any referent but must agree in gender and number with the noun following it:

'Empleados cuyo jefe'
('Employees whose manager')

'Pais cuya poblacion'
('Country whose population')

'Secretaria cuyos jefes'
('Secretary whose managers')

'Naciones cuyas superficies'
('Nations whose surfaces')

In the rules for Relative Clause formation the maximum information about the referent features is picked up from the whole clause and transmitted to the resulting Noun Phrase, in order to correctly identify the referent.

When dealing with Declarative Sentences rules are written in the Spanish Grammar to allow for more than one complement to be placed to the left of a verb; this is intended for Relative Clauses like:

'Paises a los que Italia exporta vino'

'Productos que Italia exporta a Alemania'

Actually, because of Spanish largely free word order, it is also possible for more than one complement to appear on the left of the main verb. However, most of these constructions sound odd or unnatural and are not expected.

n)
Verbs of type NAA (requiring Nominative and two Accusatives) and NAG (requiring Nominative, Accusative and Genitive) do not exist in Spanish, therefore the rules written in the German Grammar for these additional objects need not be provided.

o)
The rules for Coordination in Spanish follow quite closely the German model. Only the coordination of Noun Phrases has been kept less complicated probably due not to the greater simplicity of coordination in Spanish but to the smaller sophistication of the balance checks in the coordinated structures. These rules will certainly have to be revised.

On a coordinated Noun Phrase the plural feature is always set, the singular feature only in case the conjunction is 'o' ('or') and at least one of the constituents is singular, because in this case the verb can agree with its nearest Noun Phrase, or with the coordinated Noun Phrase. The masculine feature is set if one of the Noun Phrases are feminine.

Special rules are needed for coordination with the conjunction 'sino' ('but'), because the first element of the coordination must be negated, e.g.: 'no jefes sino empleados', 'no 5000 sino 6000'.

(In some cases of singular nouns conjuncted by 'y', especially if there is no article preceding the second noun, the resulting conjuncted Noun Phrase can also be singular: 'la entrada y salida de aviones se suspendio'. As these case are very particular they have not been considered).

p)
As to Personal Pronouns some of them can only occupy special places in sentences, and this must be accounted for: the personal pronouns 'el', 'ella, 'ellos', and 'ellas' ('he', 'she', and masculine and feminine 'they', respectively) function as normal Noun Phrases; they can be subject of sentences, and become accusative or dative if preceded by a preposition. But the Personal Pronouns 'lo', 'la', 'los', 'las' (accusative), and 'le', 'se' (accusative and dative) can only be placed to the left of the verb, and immediatly preceding it. However, all the rules dealing with Personal Pronouns only describe their syntax, no interpretation routines are provided by the system .

Exemples:

'Quien le necesita?'
('Who needs him?')

'Que pais lo exporta (a Alemania)?'
('Which country exports it (to Germany)?')

'Quien le vende (un auto)?'
('Who sells him (a car)?')

'Quien se lo vende?'
('Whow sells it to him?')

And sometimes a Personal Pronoun is used as a redudant dative to emphasize or further explain the indirect object of the sentence:

'Quien (le) vende autos a Italia?'
('Who sells cars to Italy?')

'(Le) suministra Italia vino a Alemania?'
('Does Italy supply Germany with wine?)

Rules are provided to describe this particularity.

4. CONCLUSIONS

An overview of the grammar accepted by the User Specialty Languages parser has been given. The main lines of the Spanish Grammar have been described, together with its particularities in relation with the German one upon which it is based.

It has been shown that there are many small differences and minor details changing from one grammar to the other, but the main lines have been kept, and what is more important, the interpretation routines needed for the semantic part of the USL System have been also used in the Spanish version almost unchanged. This indicates that this same approach can apply to other languages (as has already been done for example with English), and just by writing a grammar in the USL format a whole natural language interface to a relational database could be obtained. The system is ready for use in real application environments, and user experiments and studies have even been performed with the original German version (2,3). The Spanish version is less developed and tested, and needs further revisions and improvements, but we hope it will soon reach the same degree of applicability of its German counterpart.

REFERENCES

1) Astrahan M.M., et. al.
'System R: Relational Approach to Database Management'
ACM Trans. on Database Systems, vol.1, nº 2, June 1976.

2) Kettler W., Schmidt A., Zoeppritz M.
'Erfahrungen mit zwei natuerlich-sprachlichen abfragesystemen', TR 81.01.001, IBM Germany, Heidelberg Scientific Center, 1.981.

3) Lehmann H., Ott N., Zoeppritz M.
'User experiments with Natural Language for Data Base access'
Proceedings 7 th. International Conference on Computational Linguisitics, Bergen, 1.978.

4) Lehmann H.
'Interpretation of Natural Language in an Information System', IBM J. of Research and Development, vol. 22, nº 5, september 1.978.

5) Ott N., M. Zoeppritz
' USL - An Experimental Information System Based on Natural Language' in L Bolc (ed.): Natural Communication with Computers, vol. 2, Carl Hanser Verlag, Muenchen-Wien, 1.979.

6) Lehmann H.
'A System for Answering Questions in German'
Proc. 6th. International ALLC Symposium,Cambridge - England, 1.980.

7) Real Academia Española
'Esbozo de una nueva gramatica de la Lengua española'
Espasa-Calpe, Madrid, 1.979.

8) Sopeña L.
'Grammar of Spanish for User Specialty Languages'
TR 82.05.004, IBM Germany, Heidelberg Scientific Center, 1.982.

9) Zoeppritz M.
' Syntax for German in the User Specialty Languages System'
forthcoming.