THE MULTILINGUAL ENTITY TASK
A DESCRIPTIVE ANALYSIS OF ENAMEX IN SPANISH

Don D. Anderson
Department of Defense
Fort George G. Meade, MD 20755-6000
dondande@romulus.ncsc.mil
(301) 688-6149

1. Introduction. The task involved identifying and typing all named entity expressions (ENAMEX), numerical entity expressions (NUMEX), and temporal entity expressions (TIMEX) in Spanish news articles. The analysis of the data suggests that focusing on the high frequency expressions results in a higher payoff. This report looks primarily at ENAMEX expressions because they accounted for nearly three-quarters of the taggable data, and because they involved some of the more difficult tagging decisions.

Since ENAMEX accounts for so many of the taggable expressions, it makes sense that one can maximize performance by perfecting ENAMEX identification. A sample analysis of test results for recall (REC) illustrates how this works. REC, in this context, is the number of correct tags divided by the number of possible tags. In the test results, the average REC for NUMEX expressions was 91, that for TIMEX expressions 94, while that for ENAMEX was 88. These scores represent the average of all the participants' scores for correct tags divided by the total possible tags. Similarly, the total overall REC for all taggable expressions -- 89.6 -- would be found by dividing the total correct tags -- 2,977 -- by the total possible tags -- 3,320. On the other hand, if ENAMEX identification were improved enough to obtain a REC of 93 with no change in NUMEX and TIMEX RECs, then the overall average REC would jump to 94.8, an increase of roughly 5 percentage points. The example shows that a relatively small improvement in an area that accounts for a large percentage of the data (73%+ in the case of ENAMEX for Spanish) will result in a relatively large overall improvement. If the improvement had been only in NUMEX (which accounts for only about 9% of the overall data), for example, the overall improvement would have been almost negligible.

2. LOCATION data. ENAMEX expressions consisted of three types: LOCATION, PERSON, and ORGANIZATION. The analysis of the Spanish data showed that close to half (43%) of ENAMEX were LOCATION. A closer, more detailed, look at the LOCATION data suggests that the high payoff indicated by the average REC and precision (PRE) scores was achieved because most of the data were listable. Over three-quarters (76%) of the LOCATION type entities were country aggregates, countries, or capital cities. The list of these names is fairly small -- around 400 -- and is comprised of the 187 independent country names with their capital cities plus about 20-30 country aggregate names. States, provinces, and major cities accounted for 9% of the LOCATION data. When these names are added, the list is still relatively small but now accounts for about 85% of the LOCATION data in the corpus. Some of the remaining 15% of the LOCATION data is patternable as it contains key words such as *cerro* 'hill', *río* 'river', provincia 'province', and *lago* 'lake'.

The following examples are fairly representative of LOCATION data:

- Aggregates, countries, capital cities
  (1) *América Latina* 'Latin America'
  (2) *Alemania* 'Germany'
  (3) *Moscú* 'Moscow'
- States, provinces, major cities
  (1) *Nueva York* 'New York'
  (2) *la provincia de Río Negro* 'the province of Black River
  (3) *Ginebra* 'Geneva'
- Other
  (1) *la banda de Gaza* 'the Gaza strip'
  (2) *la Casa Blanca* 'the White House'
  (3) *el cerro San Cristóbal* 'San Cristóbal hill'

It is interesting to note that a simple list of the Spanish equivalents of the country aggregates, countries, and

major cities is not sufficient to account for all the listable LOCATION data. The news articles also contained country abbreviations such as *EEUU* (*Estados Unidos* 'United States'), *EAU* (*Emiratos Arabes Unidos* 'United Arab Emirates'), and *GB* (*Gran Bretaña* 'Great Britain')

3. PERSON data. High REC and PRE scores were achieved in this area beecause the data were either listable or patternable. Since the corpus is made up of news articles, it is not surprising that chiefs of state and members of governent account for 37% of the PERSON data. The other categories include known figures from fields such as sports, entertainment, the military, and religion.

Although less than 10% of PERSON names appeared with a title, around 30% were preceded or followed by an explanatory phrase regarding the profession or nationality of the person named. Thus, to a certain extent they are patternable. A large majority of names, on the other hand, must be identified by means of a list. As would be expected, many of these names will include alternate short forms. Thus, for example, all references to *Salmon Rushdie* after the first one are simply *Rushdie* in the Spanish news article about the novelist.

The following are representative examples of PERSON names in the corpus:

- Chiefs of state and members of governement
  (1) *el presidente francés Jacques Chirac*
  'the French president Jacques Chirac'
  (2) *el presidente uruguayo Luis Alberto Lacalle*
  'the Uruguayan president Luis Alberto Lacalle'
  (3) *el Secretario de Trabajo Robert Reich*
  'the Secretary of Labor Robert Reich'

- Other names
  (1) *el escritor británico de origen indio Salmon Rushdie*
  'the British writor of Indian origin Salmon Rushdie'
  (2) *el líder rebelde angoleño Jonas Savimbi*
  'the Angolan rebel leader Jonas Savimbi'
  (3) *la actriz francesa Carole Bouquet*
  'the French actress Carole Bouquet'

4. ORGANIZATION data. Acronyms or acronym-like names accounted for 57% of the ORGANIZATION data. Another 27% occurred with some kind of organizational designator and were therefore patternable, while the remaining 16% were listable. The average REC and PRE scores for ORGANIZATION data were about 8 points lower than those for PERSON and LOCATION. While these scores do not necessarily mean that the systems had problems with acronyms, it seems logical to start focusing development energies here because acronyms account for such a high percentage of the data.

At first glance, it appears that the high percentage of acronyms coupled with their complexity will present a real challenge to the developer.

The problem with acronyms and acronym-like names is how to determine whether or not a word is an acronym. Acronyms in the Spanish data ressemble English acronyms and are of at least three different types: (1) the first letters of key words in the ORGANIZATION name (such as *AFA*, the *Asociación del Futból Argentino*); (2) the initial syllables of key words or a mixture of first letters and initial syllables (such as *MINUHA, Misión de la ONU en Haití* 'United Nations mission in Haiti'); or (3) ad hoc (such as *G7*, where the *G* stands for *grupo* and *7* for *siete* in *grupo de siete países más industrializados*

Compounding the problem are foreign loan acronyms such as NLRB (National Labor Relations Board) and acronyms that are never expanded in the article -- such as AFP (Agence France Presse).

Closer analysis shows that virtually all acronyms are of the ORGANIZATION type. Furthermore, nearly all acronyms are in uppercase and so are fairly easy to identify. Abbreviations such as *EEUU*, *EAU*, and *GB* mentioned previously can be filtered out and once the acronym is identified, an attempt can be made to match it to an occurrence in the all uppercase header area common to all the news articles.

Over a quarter of the ORGANIZATION names are patternable to a degree since they include an organizational designator such as *banco* 'bank', *grupo* 'group', *organización* 'organization', and *asociación* 'association'. The problem is that even though the majority of ORGANIZATION names with designators occurred with the designator in initial position, it is not clear how to identify where the organizational name ends. The following examples will illustrate:

(1) *Banco Mundial*
(2) *banco First Albany*
(3) *Comité de Familiares de Detenidos Desaparecidos*

(4) *Comité nacional de la lucha contra las epidemias de Kikwit*

(5) *Organización de Cooperación y Desarrollo Económico*

(6)*Organización Mundial de la Salud*

5. Summary. In summary, the analysis of the Spanish data shows that as a general rule, most of the ENAMEX occurrences are listable or easily identifiable by non-linguistic means. Identification of patternable ENAMEX is, on the other hand, not straightforward due to the difficulty of determining precise patterns. Comparison of this analysis with average REC and PRE scores suggests that there is a correlation between high scores and a high frequency of occurrence of data that are mostly listable or easily determined by means of well-defined patterns.