

Featuring of Sex-Dependent Nouns in Databases Oriented to European Languages*

Igor A. Bolshakov¹ and Sofia N. Galicia-Haro²

¹ Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico

igor@cic.ipn.mx

² Faculty of Sciences
National Autonomous University of Mexico (UNAM)

Mexico City, Mexico
sngh@ciencias.unam.mx

Abstract. It is argued that human-denoting nouns in European languages forming pairs like English *steward* vs. *stewardess*, or Spanish *jefe* vs. *jefa* ‘chief’, or German *Student* vs. *Studentin* ‘student’, or Russian *moskvič* vs. *moskvička* ‘Muscovite’ may be featured in factographic databases conjointly as Sex-Dependent Nouns? a special part of speech. Each SDN has two forms, maybe coinciding, selected when necessary by the sex of the denoted person. SDN notion ensures a kind of universality for translation between various languages, being especially convenient in languages with gender of nouns implied by sex. We base our reasoning on Spanish, French, Russian, German, and English examples.

1 Introduction

European languages have numerous pairs of human-denoting nouns that are the same in meaning but differ in the sex of their bearer. The semantics of these pairs covers at least the following groups:

1. Profession, occupation or official position (in English: *steward* vs. *stewardess*; in Spanish: *jefe* vs. *jefa* ‘chief’; in Russian: *učitel* vs. *učitel’nica* ‘teacher’; in German: *Student* vs. *Studentin* ‘student’; in French: *directeur* vs. *directrice* ‘director’);
2. Nationality or confession (in English: *Frenchman* vs. *Frenchwoman*; in Spanish: *católico* vs. *católica* ‘Catholic’; in Russian: *nemec* vs. *nemka* ‘German’; in French: *suédois* vs. *suédoise* ‘Swede’);
3. Dwelling locality (in French: *parisien* vs. *parisienne* ‘Parisian’; in Spanish: *madrileño* vs. *madrileña* ‘dweller of Madrid’; in Russian: *sibirjak* vs. *sibirjačka* ‘Siberian’; in French: *sévillan* vs. *sévillane* ‘Sevillian’);
4. Main feature of personality (in English: *madman* vs. *madwoman*; in Spanish: *histérico* vs. *histérica* ‘hysterical’; in Russian: *durak* vs. *dura* ‘stupid’; in French: *délinquant* vs. *délinquante* ‘delinquent’).

* Work done under partial support of Mexican government (CONACyT, SNI, SIP-IPN). Many thanks to Steve Legrand for the proofreading of the manuscript.

Because of the same meaning, the pair counterparts are considered synonymous. Morphologically, they usually have the same root, and the repertoire of sex-forming suffixes is rather limited (about a dozen suffix pairs, depending on language).

In contrast to the pairs, there exist many nouns of the mentioned groups whose dependence of sex is not expressed explicitly. The same form is used, but their co-reference relationships in context are expressed by sex-dependent personal pronouns, e.g., *he is a teacher* vs. *she is a teacher*.

Meanwhile, in languages with grammatical gender (e.g., of Romance or Slavic families) such nouns are of masculine or feminine gender for males and females respectively. The sex as a semantic feature is expressed by means of the morpho-syntactic category of gender of dependent articles (Sp. *el estudiante*_{MALE} vs. *la estudiante*_{FEMALE} ‘the student’), dependent adjectives (Sp. *nuevo estudiante*_{MALE} vs. *nueva estudiante*_{FEMALE} ‘new student’) or syntactically related predicatives (Sp. *estudiante*_{MALE} *está cansado* vs. *estudiante*_{FEMALE} *está cansada* ‘student is tired’; Rus. *vrač*_{MALE} *skazal* vs. *vrač*_{FEMALE} *skazala* ‘physician said’).

The Sex-Dependent Nouns are already described in the works [1, 2, 3] of general linguistics. In [3] two different viewpoints are compared. One of them considers SDN pairs as members of the same lexeme, and this is usual for many modern Spanish and French dictionaries. E.g., the French pair *fou* vs. *folle* ‘insane’ appears in the same dictionary entry as inflectional forms of the same entity. Hence it is admitted that such nouns get inflectional forms of each sex according to the specific gender?the way adjectives or participles do.

The contrary viewpoint considers *fou* and *folle* as two different lexemes linked by a derivational relation. Hence, the sex is included in the semantic description of each lexeme. Russian lexicography always prefers this viewpoint, whereas Romance language lexicographies adopt it only in cases when the difference between the two forms is rather significant, e.g. for *actor/actriz* ‘actor/actress’ in Spanish.

In this paper, we argue for the first viewpoint and suggest the introduction of a new part of speech, i.e., Sex-Dependent Noun (SDNs), for use in computational linguistics. SDNs are especially applicable to languages with gender. In these languages, they are similar to adjectives, the number and gender of which have the same word-forming characteristics. Both number and gender of SDNs are semantically loaded: they reflect the number of persons and their sex. Hence, they both should be recognizable when analyzed as a meaning unit in the semantic representation of the analyzed text. In synthesis, the sex feature is used for the selection of target word form from the pair (if the forms are different) and for correct agreement of words referentially related with the target.

We base our considerations on Spanish, French, Russian, German, and English languages, revealing few differences in properties of SDNs in various languages.

2 Sex-Dependent Nouns in Spanish

The number of SDNs in Spanish is tremendously high, i.e. several thousands. The Anaya dictionary contains ca. 3600 nouns of this type, i.e. 18% of the total nouns. The most common morpho-classes of SDN formation are shown in Table 1 (\emptyset is empty string).

The most common are the morpho-classes 1 to 3, both broadly spread and actually productive. Indeed, the modern Spanish easily admits new female forms for profession