A proposal for a multilevel linguistic representation of Spanish personal names

Orsolya Vincze
Universidade da Coruña
ovincze@udc.es

Margarita Alonso Ramos
Universidade da Coruña
lxalonso@udc.es

Abstract

This paper proposes a multilevel representation of personal names, with the aim of offering an economical treatment for these expressions, which makes a clear distinction between ontological information, described in a *name database,* and linguistic levels of representation. Adopting the linguistic model and formalisms provided within the Meaning⇔Text framework (Mel'čuk 1988), it is argued that, contrary to other proper names (e.g. organizations, toponyms, etc.), which should be treated similarly to idioms, complex personal names such as *José Luis Rodríguez Zapatero* should not be represented as single units at any linguistic level nor in the lexicon. Variant forms referring to a concrete person (e.g. *José Luis Rodríguez Zapatero, Rodríguez Zapatero, Zapatero, Z.P.*) are accounted for by a set of rules connecting the *name database* and the semantic level.

1 Introduction

Proper names have traditionally occupied a rather marginal position in linguistic description. As a consequence, the systematic and formalized description of their syntactic and morphological behavior remains largely unattended. More recently, in the field of natural language processing (NLP), the treatment of proper names has been put into focus, as a consequence of the growing interest in tasks involving the recognition of *named entities*, a set of expressions characterized by having a unique reference (e.g. Vitas et al. 2007).

A problem going further than the mere identification of segments of texts as proper names, which is generally solved using simple heuristics (cf. Krstev et al. 2005: 116), is that of the treatment of the various ways a particular entity can be referred to (Nadeau and Sekine 2007). For instance, in journalistic texts, the current Spanish prime minister can be designated by either one of the following strings: *José Luis Rodríguez Zapatero, Zapatero,* or *Z.P.* It has been found that NLP applications dealing with this latter, more complex question can profit from information on the linguistic properties of names (e.g. Charniak 2001; Gaizauskas et al. 2005; Vitas et al. 2007). One way of tackling the problem, proposed by the authors of *Prolexbase* (Krstev et al. 2005), a multilingual ontology of proper names, is that of explicitly listing variant forms of names in a lexical database.

The aim of the present paper is to propose a representation of Spanish personal names, wherein variant forms can be treated in a more economical way. For this, we have adopted the linguistic model proposed within the Meaning⇔Text framework (MTT, Mel'čuk 1988). To our knowledge, no attempt has been made to formally integrate personal names in any such comprehensive linguistic model, therefore, this proposal should be considered as rather tentative.

The most important feature of our description is that we suggest a clear distinction between ontological information, contained in the *person database,* where a person is conceived as a single entity, and linguistic representation, where personal name strings are analyzed as complex structures constituted by name elements. Consequently, as we will show, variant forms can be accounted for by a set of rules establishing correspondences between the person database and the linguistic levels of representation.

Note that, in what follows, we will use the more generic term *proper name* to refer to those expressions which constitute the names of geographical locations, organizations, institutions, persons, etc., while the more specific term *personal name* will be used for the expressions that name particular individuals.

2 Related work

2.1 Encyclopedic vs. linguistic description of proper names

The definition of the notion of proper names has been formulated in various ways in linguistics, mainly proposing an opposition of this class to that of common nouns on the basis of their different semantic and/or referential properties. We do not intend to discuss this issue in detail; however, it is relevant to note that the existence of such an obvious difference lies at the root of the lexicographical tradition of excluding proper names from dictionaries, and transferring them to encyclopedias (Marconi 1990). This practice has been challenged by some authors, (e.g. Lázaro Carreter 1973; Mufwene 1988; Higgins 1997) arguing that, whatever the content of these expressions, their linguistic properties, such as gender, number, pronunciation, variant spellings, etc. should be described systematically.

Concentrating on the case of personal names, we find that, like other proper names, these are generally excluded from dictionaries; that is, we will not find dictionary entries with names of specific persons, given that this information is considered to belong to the encyclopedia. More importantly, name elements such as given names like *José*, their non-standard diminutive form *Pepe*, and surnames like *Rodríguez* are also excluded from the lexicographical tradition. Nevertheless, we do find some cases of derived relational adjectives that make reference to specific persons, e.g. *Freudian* with reference to Sigmund Freud. This latter aspect has been pointed out by, for instance, Lázaro Carreter (1973) and Higgins (1998), who claim that it violates the self-sufficiency principle in lexicography, namely, definitions of these adjectives point to entities – specific persons – on whom often no information is provided in the dictionary.

Within the field of NLP, it is claimed that named entity recognition systems are able to function quite efficiently on the basis of simple heuristics (Krstev et al. 2005: 116). This may be the reason why researchers working in this field are generally not concerned with describing specific linguistic properties of these expressions in a systematic way. Although lexical resources such as ontologies or knowledge-based systems are created for named entity tasks (e.g. Morarescu and Harabagiu 2004; Rahman and Evens 2000), these are generally applied for the semantic classification of named entities. In consequence, they are merely designed to incorporate encyclopedic information in a formal, computerized lexicon, leaving linguistic properties of proper names unattended.

On the contrary, the description of the linguistic properties, together with the formal and orthographic variants of proper names, seems to be rather important in the case of more complex tasks such as identifying *aliases*, that is, the various ways an entity can be spelled out in a text (cf. Nandeau and Sekine 2007: 16), or for computer-assisted translation and multilingual alignment (Maurel 2008). For instance, as illustrated in (1), a person, such as *Sigmund Freud* can be referred to by variant name forms, as well as by a derived relational adjective. Moreover, some languages may prefer one formulation to the other, and a language may completely lack a particular derivative form (Vitas et al. 2007: 119).

(1) *Sigmund Freud's/S. Freud's/Freud's/the Freudian theory of human personality*

Prolexbase (Krstev et al. 2005; Maurel 2008, etc.), a multilingual relational database of proper names has been created with the aim of proposing a solution for the problem posed by variant forms of proper names. Consequently, besides conceptual or encyclopedic information, it also contains description of formal variants. Each entity is represented by a single, language independent node, which is linked to a lemma in each specific language, representing the base form of the given proper noun, which is in turn specified for all of its variant forms. For example, as shown in Figure 1, the same ID is associated with the French and the English lemmas, *États-Unis* and *United States* respectively, and the latter is specified for its variant realizations *United States of America, US, USA, as* well as the adjective *American*.

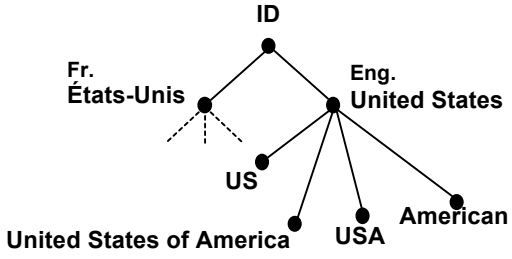


Figure 1: Representation of different forms of proper names in Prolexbase (adapted from Maurel 2008: 335)

2.2 Representation of the structure of personal names in syntactically annotated corpora

The syntactic representation of personal names and of proper names in general, to our knowledge, has not received sufficient attention. In descriptive grammars, authors tend to limit their analysis of the structure of these expressions to proposing a classification based on their lexical elements: for instance, many proper names are composed only of word forms that can be classified as proper names themselves (e.g. *Socrates, Switzerland*), while others are more similar in their structure to regular noun phrases (e.g. *United Kingdom, University of Cambridge*), given that they contain adjectives and common nouns (e.g. Quirk et al. 1985: 288-294; Allerton 1987: 67-69). At the same time, after a brief look at syntactically annotated corpora, we arrive at the conclusion that within the field of NLP, there is no consensus on whether to analyse the syntactic structure of names. Namely, treebanks in general differ in treating multilexemic proper names as single nodes, e.g. Spanish *AnCora Corpus* (Martí et al. 2007) and Portuguese *Floresta Sintá(c)tica* (Afonso et al 2002), or as proper subtrees, e.g. *Prague Dependency Treebank* (PDT, Hajičová et al. 1999; Böhmová et al. 2005).

As for the more specific case of the representation of the syntactic structure of personal names, the number of existing proposals is rather limited. For instance, Anderson (2003: 374) suggests that complex forms of personal names are headless compounds, whose elements, the given name and the family name are juxtaposed, given that, one may perceive the given name modifying the family name, or vice versa, depending on the context. Within the dependency framework, the PDT provides an analysis where all other elements are syntactic dependents of the rightmost element of the name string, generally the last name, and are represented as adnominal modifiers (Böhmová et al. 2005: 836). From the perspective of the MTT, Bolshakov (2002) suggests representing Spanish personal names as dependency chains where the family name depends on the given name, and proposes a specific type of surface syntactic relation, *nomination appositive*, to describe the dependencies between their components.

3 The linguistic status of personal names

Given that we aim at proposing a linguistic description for personal names, we have to raise the question of what kind of linguistic units these expressions are and how they should be represented on each level of representation proposed by the linguistic model of the MTT (e.g. Mel'čuk 1988).

An important feature of our framework is the clear split between linguistic and non-linguistic level. Following this idea we propose to describe ontological information on each entity in the *name database*, separate from the linguistic representation, attending solely to linguistic properties of name elements. In this way, we obtain a more economical treatment of variant forms of personal names via a set of rules operating between these two main levels of representation, and avoid explicitly listing variant forms of names in a lexical entry (cf. Tran and Maurel 2006:119-120).

In syntactic analysis, as in the case of some *treebanks*, proper names are often treated as idioms, that is, indecomposable chains. However, the MTT proposes a more multifaceted treatment for idioms. Within this framework the semantic unity of full idioms is reflected by representing them as a single lexical unit in the dictionary and, consequently, as a single node at the deep syntactic level, while they are assigned a proper subtree representation at the surface syntactic level, indicating their internal structure. The reason for this is not only the lack of semantic compositionality characterizing these expressions, but also the structural irregularities they present in comparison with regular, compositional expressions.

We would like to underline that, from our point of view, an important distinction should be made between the representation of names of organizations, toponyms, etc., on the one hand, and personal names, on the other hand. We claim that expressions belonging to the first group (e.g. *Organization of United Nations*) should be treated similarly to full idioms, attending to semantic non-compositionality. In contrast, we suggest complex personal names to be represented as complex structures at all linguistic levels: as various lexical units in the dictionary, as a compound lexical node at the deep syntactic level, that is, a lexeme constructed from various actual lexemes, similarly to the case of com-

pound lexemes (Mel'čuk forthcoming), and as a tree at the surface syntactic level.

This proposal is based, on the one hand, on the assumption that the structure of personal names can be considered as regular, that is, it can be sufficiently accounted for by a specialized mini-grammar. On the other hand, we claim that, contrary to full idioms which cannot be analysed in terms of the meanings of their components, in the case of names, the meaning of each element, that is, the meaning of each given name and each family name, can be represented as an independent *denomination predicate*, e.g. *José = person X is called José*. We have adopted this concept from Gary-Prieur (1994), according to whom the *meaning* of a proper name is distinct from its *content* defined as a set of properties attributed to the referent.

We assume that the possibility of referring to a person by variant name forms suggests that name elements retain their meaning and can have the same referential content whether used as a full string or independently (as in 2a). Thus, as we show in sentence (2b), meanings of name elements seem to be autonomous within a name string, which is further demonstrated by the fact that they are accessible for coordination (see 2c). Finally, we consider that utterances like (2d) and (2e) can be considered potential cognitive, or, more precisely, referential paraphrases (cf. Milićević 141-145).

(2a) *That was the first time I met María Lamas, although I'd known María's sister for a long time*.
(2b) *The author Manuel Rivas is called the same as your father (who is called Manuel González*).
(2c) *Ana and María Lamas/Ana Lamas and María Lamas are visiting us this week.*
(2d) *María Lamas*
(2e) *the girl whose name is María and whose surname is Lamas*

4. Linguistic representation of Spanish personal names

As we have said, our proposal distinguishes two main levels of description: the person database and the linguistic representation. As for the linguistic description, in accordance with the MTT framework, we foresee a *dictionary*, where name elements, that is, both given names and family names are listed and specified for their linguistic properties. Furthermore, we deal with the following three levels of linguistic representation: *semantic representation* (SemR), the *deep syntactic* (DSyntR) and the *surface syntactic representations* (SSyntR). Each two consecutive levels are connected by a set of rules that serve to establish correspondences between them. Among these, we will limit ourselves to those operating between the person database and the semantic level.

For the purpose of the present paper, we will limit ourselves to the analysis of the most common forms of personal names in European Spanish, which, in their full form, consist of a number of given names, followed by two family names, e.g. *José Luis Rodríguez Zapatero*. Note that full forms of Spanish names usually contain two family names, the first of these being the first family name of the father and the second the first family name of the mother.

4.1 The person database

The *person database* contains a list of all individuals relevant in a given application. Naturally, it would be impossible to compile a list of all human beings in the world, so, for practical purposes, the content of this component will always have to correspond to specific research objectives. For each individual, several name attributes are specified, such as a) first family name, b) second family name, c) first given name, d) second given name, e) nickname, and f) derived lexical units. Sometimes an individual can be referred to by different full names depending on the context, in these cases, attributes have to be specified under such fields as *birth name*, *public name*, *maiden name*, etc. (Cf. Bartkus et al. 2007). See Figure 2 for an example of the representation corresponding to *José Luis Rodríguez Zapatero*.

ID=134567
First family name=Rodríguez
Second family name=Zapatero
First given name=José
Second given name=Luis
Nickname=Z.P.
Derivate = zapateriano

Figure 2: Representation in the person DB

At this level, the attribute *nickname* refers to a form that is used to name a particular individual. This form does not correspond to

standard nicknames or diminutives (see section 3.2), which can make reference to any individual carrying a particular name. Likewise, as we have already explained, derivative forms included in the ontological representation also make reference to a specific person e.g. *freudiano → Sigmund Freud; cervantino → Miguel de Cervantes, isabelino → Queen Elizabeth I and Queen Elizabeth II of Spain, Queen Elizabeth I of England.*

The name database should also include relevant extralinguistic or encyclopedic information on each individual. This information may have certain importance in the identification of a name as referring to a specific person on the basis of context, for instance, appositives like *presidente, general, secretario, director,* etc. (cf. Arévalo et al. 2002). As we have seen, encyclopedias and certain resources developed for NLP applications generally concentrate on this kind of information. However, since our purpose is to look at personal names from a strictly linguistic point of view, we won't discuss this aspect in more detail.

4.2 The dictionary

The *dictionary* should include a complete list of name elements, that is, given names and family names together with their variant and derivative forms. This implies that our formal dictionary does not include the full form of the name, and hence, encyclopedic information on a specific person, e.g. *José Luis Rodríguez Zapatero*, instead, it specifies the following information (see Figure 3).

José:	proper name, given name, masculine
	Nickname: *Pepe*
Luis:	proper name, given name, masculine
Rodríguez:	proper name, family name, weak
Pepe:	[=**Nickname**(*José*)] proper name, nickname, masculine
Zapateriano:	adjective, related to *ID134567*
Zapatero:	proper name, family name
Z.P.:	nickname for *ID134567*

Figure 3: Representation in the dictionary

Note that in the case of each name element, we include information on syntactic class (*proper name*) and specify the subcategory (*given name* or *family name*). We consider the latter distinction necessary, given that, as we will show later, we perceive a difference in the syntactic combinability of these classes[1].

Lexical entries of given names indicate irregularly derived standard nicknames. For instance, in the case of *José*, we include the form *Pepe* but not regularly derived *Josito*[2]. These variant forms also receive their own dictionary entry, while derived forms or non-standard nicknames, like *Zapateriano* or *Z.P.,* constitute an individual entry, without any link to the base form. Note that, as we have already discussed, these forms make reference to a specific person, instead of e.g. all persons called *Zapatero*, that is why, their reference is specified via an ID, assigned to the person in the person database.

Another property of both given- and family names that we find important from the point of view of lexical description, is the feature of *weakness*. In the case of female compound given names such as *María Teresa, María Dolores*, etc. Spanish speakers will generally opt for using the second element, contrary to other compound cases like *Fernando Manuel* or *Rosa María*, where generally the second given name is omitted. Similarly, in the case of family names, there is a preference towards retaining the second family name when it is perceived as more salient. An example would be the case of the Spanish president *José Luis Rodríguez Zapatero*, commonly referred to as *Zapatero* and not as *Rodríguez*. In both cases, the attribute *weakness* seems to be related to the frequency of use of these name elements, however, further empirical research would be needed to establish clear criteria. For some frequency information on compound given names, see (Albaigès 1995: 82-83).

Finally, we find worth mentioning that there are certain forms of given names for which it may be problematic to decide whether they should be treated as compounds containing two independent name elements or they

[1] Naturally, the choice of one or another combination of these name elements to refer to an individual also reflects pragmatic, sociolinguistic, etc. differences, factors which are beyond the scope of this study.

[2] Note that the distinction between regularly and not regularly derived standard nicknames may not be as straightforward as it may seem at first sight. Spanish given names generally, but not always, receive the diminutive ending *–ito/a* as in *Miguel→Miguelito, Rosa→Rosita,* but *Carlos→Carlitos,* and not **Carlosito*; *Mercedes→Merceditas,* and not **Mercedesita*. (We would like to thank one of the anonymous reviewers for pointing this out.)

should be stored as a single lexical unit. For instance, in the forms *María del Carmen, María del Pilar,* etc., similarly to cases we have just seen, *María* tends to behave as a weak element, however, the second part *del Pilar* or *del Carmen* is not autonomous, e.g. *María del Carmen Álvarez/Carmen Álvarez/*del Carmen Álvarez*. Furthermore, certain compounds correspond to a single diminutive form, e.g. *María del Carmen=Maricarmen/Mari Carmen*, *José Miguel=Josemi, José María=Chema, María Jesús=Chus,* while others, like *José Luis* or *Miguel Angel*, although they do not have a corresponding single diminutive form, are often perceived as a single word form.

4.3 Semantic representation (SemR)

As we have already suggested, in formulating the SemR, we have adopted the concept of *denomination predicate,* (Gary-Prieur 1994) to represent the *meaning* of names. Consequently, we conceive of each name element as including a predicate, e.g. *José = person X is called José* so that the representation of the sequence used to refer to a specific person called *José Luis Rodríguez Zapatero* would be as in Figure 4.

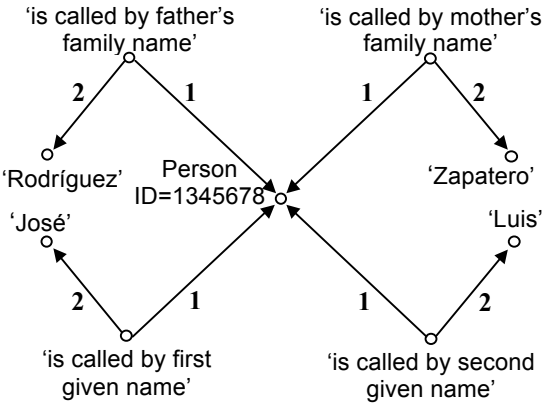


Figure 4: SemR of the name *José Luis Rodríguez Zapatero*

As shown in Figure 4, in the cases where more than one name element of the same category (i.e. given name or family name) is used, the semantic representation is enriched with more specific information. For instance, full forms of Spanish names usually contain two family names, as we have said, the first of these coming from the father and the second from the mother. When only one name element of a category is used, this information would not necessarily be present in SemR. As a consequence, simpler semantemes could be used, e.g. if the current Spanish president is referred to by the form *Zapatero,* the semanteme 'family name' instead of 'mother's family name' would be used in the SemR.

4.4 Deep- and surface syntactic representation (DSyntR and SSyntR)

The syntactic representation of personal names has not been studied in detail within the Meaning⇔Text framework, the only proposal we know about being that of Bolshakov (2002).

We propose representing personal names on the DSynt by a single node, in a similar way as compound lexemes are represented. As pointed out by Mel'čuk (forthcoming), compound lexical units that are fully compositional potential lexical units do not need a lexical entry in the dictionary, given that they are constructed in a regular way through the combination of full lexemes. Their internal structure is considered a syntactic island for DSyntS rules, but it is specified as a tree-like structure whose nodes are labelled with the compounding lexemes, in order to provide information for the linear ordering of components. In a similar way, we propose representing personal names as potential lexical units constructed out of element names, see (3a) and (3b).

(3a)

- [JOSÉ→LUIS→RODRÍGUEZ→ZAPATERO]

(3b)

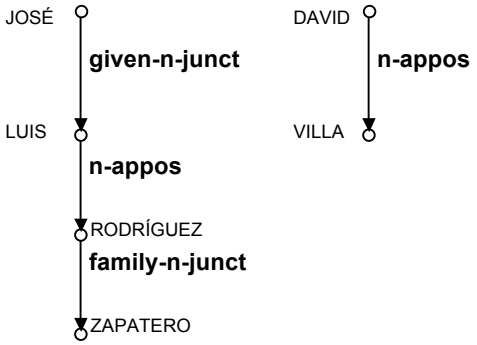
- [DAVID→VILLA]

However, on the SSynt level, personal names will be represented as proper sub-trees, in the same way as idioms, following Mel'čuk (1988 and 2009). Nevertheless, we have found that the special characteristics of personal names do not lend themselves easily to determining the directionality of syntactic relations on the basis of the criteria proposed by Mel'čuk (2009). As a consequence, we have decided to adopt Boshakov's (2002) scheme, where, as we have already mentioned, name elements form a dependency chain headed by the first given name. Considering the lack of other criteria, we believe this kind of representation to be convenient, given that it facilitates linearization, contrary to, for instance, PDT type representation (see section 2.2).

For labelling dependencies, we have decided to introduce three different syntactic relation types to represent relations between the name elements that concern us, that is, given names and family names. Our decision was

based on one of the criteria provided by Mel'čuk (2009: 34-35), namely that every relation type has to have its prototypical dependent, which can be substituted for the actual dependent in any configuration, resulting in a correct structure. Consequently, we propose *name appositive* to represent the relation between the last given name or a standard nickname and the first family name, *given name junctive* will stand between any two given names and, finally, *family name junctive* connects the two family names, see (4a) and (4b).

(4a) (4b)

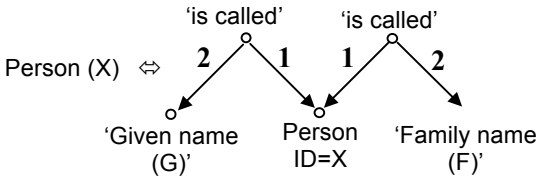


4.5 Mapping between the *person database* and the semantic level

In the MTT framework correspondences between two consecutive levels of linguistic representations are established by a set of rules. Similarly, we propose a series of rules for mapping between the *person database* and the semantic level of our model, with the aim of providing a systematic account for the formal variants of personal names referring to the same individual. These rules reflect all possible combinations of the name elements.

By way of illustration, we will discuss the case of the complex name form consisting of one single given name and one family name[3]. For the mapping rules applied in this case see Figure 5. G1 and G2 stand for the forms filling the first and second given name attribute respectively, and F1 and F2 are the forms filling the father's and the mother's family name attribute respectively. Note that in the semantic representation, as we have discussed, a proper sub-network will correspond to each selected attribute.

Person Database ⇔ **SemR**



Conditions:

(X.G1 and not X.G2)	// Given name=X.G1
or (X.G1.weak=no)	// Given name=X.G1
or (X.G1.weak=yes and X.G2);	// Given name=X.G2
(X.F1.weak=no)	// Family name=X.F1
or (X.F1.weak=yes and X.F2.weak=yes)	// Family name=X.F1

Figure 5: Mapping Rule between the *name database* and the semantic level

We assume that on the basis of these rules and making use of both types of information stored in the *name database* and the *dictionary*, correct forms agreeing with the first name+given name pattern (G F) can be generated. For instance, for a person whose corresponding attributes are G1=*María,* G2=*Teresa,* F1=*Álvarez,* F2=*Fernandez*, we can generate the form *Teresa Álvarez*, given that name elements *María, Álvarez* and *Fernández* are specified as [+weak] in the dictionary. Similarly, these rules can serve to associate the form *Teresa Álvarez* with persons with matching attributes in the *name database*.

Note that the name of current Spanish president José Luis Rodríguez Zapatero is generally not to be used with this pattern, since first [+weak] family names followed by a [-weak] family name are rarely, and second family names are never used alone. That is, for any Spanish speaker it would result rather strange to refer to the current prime minister as *José Luis Rodriguez* and they would never refer to him as *José Luis Zapatero*. As we have already mentioned, the compound name *José Luis* shows a particular behaviour, for now, not covered by our rules. A single element, either *José* or *Luis* is used only without family names, on the contrary, when family names are used as well, these given names tend to obligatorily appear in the compound form, which may point towards the fact that this form should be treated as a single word form.

[3] Other possible variant patterns are: 1) Given name+Given name+Family name+Family name (*José Luis Rodríguez Zapatero*); 2) Given name+Given name (*José Luis*); 3) Given name+Family name+Family name (*Federico García Lorca*), 4) Family name (*Aznar*) and 5) Non-standard nickname (*ZP*).

5 Conclusion

This paper has presented a proposal for a multilevel representation of personal names with the aim of accounting for variant combinations of name elements that can be used to refer to a specific person. We have suggested that a clear distinction is necessary between ontological information and linguistic levels of representation. Adopting the linguistic model and formalisms provided by the MTT framework, we have argued that, contrary to other proper names, such as names of organizations, toponyms, etc., which should be treated similarly to full idioms, personal names are to be represented as complex structures on all linguistic levels: as various lexical units in the dictionary, a "quasi-compound" lexical node on the deep- and as a tree on the surface syntactic level. Finally, variant forms of personal names referring to a given individual have been accounted for by a set of rules establishing correspondences between the name database, containing ontological information, and the semantic level.

Acknowledgments

This work has been supported by the Spanish Ministry of Science and Innovation and the FEDER Funds of the European Commission under the contract number FFI2008-06479-C02-01. We would also like to thank Simon Mille, Igor Mel'čuk and Leo Wanner, as well as the anonymous reviewers for their valuable remarks and comments on the previous versions of this text.

References

Afonso, S., E. Bick, R. Haber and D. Santos. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In González Rodrgíguez, M. And C. P. Suárez Araujo (eds.) *Proceedings of LREC 2002*, Paris, ELRA, 1698-1703.

Albaigès, J. 1995. *Enciclopedia de los nombres propios*. Barcelona: Planeta.

Allerton, D. J. 1987. The linguistic and sociolinguistic status of proper names. *Journal of Pragmatics* XI: 61-92.

Anderson, J. 2003. On the structure of names. *Folia Linguistica: Acta Societatis Linguisticae Europaeae* XXVII(3-4): 347-398.

Arévalo, M., X. Carreras, L. Màrquez, M. A. Martí, L. Padró and M. J. Simón. 2002. A proposal for wide-coverage Spanish named entity recognition. *Procesamiento de Lenguaje Natural*, 28: 63-80.

Bartkus, Kim, Paul Kiel and Mark Marsden. Person name: Recommendation, 2007 April 15. HR-XML Consortium. http://ns.hr-xml.org/2_5/HR-XML-2_5/CPO/PersonName.html

Böhmová, A., A. Cinková and E. Hajičová. 2005. A manual for tectogrammatical layer annotation oft he Prague Dependency Treebank. Available: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf

Bolshakov, I. 2002. Surface syntactic relations in Spanish. In Gelbukh, A. (ed.) *Proceedings of CICLing-2002,* Berlin/Heidelberg: Springer-Verlag, 210-219.

Charniak, E. 2001. Unsupervised learning of name structure from coreference data. In *NAACL*.

Gaizauskas, R., T. Wakao, K. Humphreys, H. Cunningham and Y. Wilks. 2005. Description of the LaSIE System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann.

Gary-Prieur, M-N. 1994. *Grammaire du nom propre,* Vendôme, Presses Universitaires de France.

Hajičová, E., Z. Kirschner and P. Sgall. 1999. A manual for analytical layer annotation of the Prague Dependency Treebank. Available: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf

Higgins, Worth J. 1997. Proper names exclusve of biography and geography: maintaining a lexicographic tradition. *American Speech*, 72(4): 381-394.

Krstev, C., D. Vitas, D. Maurel, M. Tran. 2005. Multilingual ontology of proper names In *Proceedings of 2nd Language & Technology Conference,* 116-119.

Lázaro Carreter, F. 1973. Pistas perdidas en el diccionario. *Boletín de la Real Academia española*, 1973, 53(199): 249-259.

Marconi, D. 1990. Dictionaries and proper names. *History of Philosophy Quarterly*, 7:1, 77-92.

Martí, M. A., M. Taulé, M. Bertran and L. Márquez. 2007. AnCora: Multilingal and multilevel annotated corpora. Available: clic.ub.edu/corpus/webfm_send/13

Maurel. D. 2008. Prolexbase: A multilingual relational lexical database of proper names. In Calzolari N. et al. (eds.) *Proceedings of LREC-2008*, Paris: ELRA, 334-338.

Mel'čuk, I. 1988. *Dependency syntax: Theory and practice*, Albany, State University of New York Press.

Mel'čuk, I. 2009. Dependency in natural language. en Mel'čuk, I. and A. Polguère (eds.), *Dependency in linguistic description*, Amsterdam/Philadelphia, John Benjamins, 1-110.

Mel'čuk, I. forthcoming. *Semantics*, Amsterdam/Philadelphia: John Benjamins.

Morarescu, P. and S. Harabagiu. 2004. NameNet: a self-improving resource for name classification. In *Proceedings of LREC 2004*.

Milićević, Jasmina. 2007. *La paraphrase: Modélisation de la pharaphrase langagière,* Bern, Peter Lang.

Mufwene, S. S. 1988. Dictionaries and proper names. *International Journal of Lexicography*, 1(3): 268-283.

Nadeau, D. and S. Sekine. 2007. A survey of *named entity* recognition and classification. *Lingvisticae Investigationes*, 30(1): 3–26.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. 1985. *A contemporary grammar of the English language,* London/New York: Longman.

Rahman M. A. and Evens, M. 2000. Retrieving knowledge for a lexical database from proper noun entries in Collins English Dictionary. In *Proceedings of MAICS-2000,* 63-67.

Tran, M. and D. Maurel. 2006. Un dictionnaire relationnel multilingüe de noms propres. In *Traitement Automatique des Langues* XLVII(3): 115-139.

Vitas, D., C. Krstev and D. Maurel. 2007. A note on the semantic and morphological properties of proper names in the Prolex Project. *Lingvisticae Investigationes*, 30(1): 115-133.