Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian

Monica Monachini[1], Nicoletta Calzolari[1], Khalid Choukri[2], Jochen Friedrich[3], Giulio Maltese[5] Michele Mammini[1], Jan Odijk[4], Marisa Ulivieri[1]

[1] ILC-CNR
Via Moruzzi 1, 56124 Pisa - Italy
monica.monachini, nicoletta.calzolari, michele.mammini, marisa.ulivieri@ilc.cnr.it

[2] ELRA-ELDA
Rue Brillat Savarin, 75013 Paris - France
choukri@elda.fr

[3] IBM Germany
Gottlieb-Daimler-Strasse 12, 68165 Mannheim - Germany
jochen@de.ibm.com

[4] Utrecht University & Nuance Communications International
Trans 10, 3512 JK Utrecht – The Netherlands
jan.odijk@let.uu.nl

[5] IBM Italy
Via Sciangai 53 - 00144 Rome – Italy
giulio.maltese@it.ibm.com

Abstract

The goal of this paper is (1) to illustrate a specific procedure for merging different monolingual lexicons, focusing on techniques for detecting and mapping equivalent lexical entries, and (2) to sketch a production model that enables one to obtain lexical resources via unification of existing data. We describe the creation of a Unified Lexicon (UL) from a common sample of the Italian PAROLE/SIMPLE/CLIPS phonological lexicon and of the Italian LCSTAR pronunciation lexicon. We expand previous experiments carried out at ILC-CNR: based on a detailed mechanism for mapping grammatical classifications of candidate UL entries, a consensual set of Unified Morphosyntactic Specifications (UMS) shared by lexica for the written and spoken areas is proposed. The impact of the UL on cross-validation issues is analysed: by looking into *conflicts*, mismatches and diverging classifications can be detected in both resources. The work presented is in line with the activities promoted by ELRA towards the development of methods for packaging new language resources by combining independently created resources, and was carried out as part of the ELRA Production Committee activities. ELRA aims to exploit the UL experience to carry out such merging activities for resources available on the ELRA catalogue in order to fulfill the users' needs.

1. Introduction

The assumption behind this paper is that development, packaging and customization of Language Resources (LRs) are critical issues in view of stimulating the Human Language Technology (HLT) community and providing different prospective end-users (both the industrial market and academic institutions) with the LRs they need. The ideal resource to fulfill users' requirements is often difficult to find; conversely, the LR landscape can offer a large number of individual resources that could potentially contain what meets users' expectations. The problem is that, since they have been created by different developers for different purposes, these resources cover different types of data and linguistic phenomena; and, what's more, the information can be expressed in diverging formats. In this scenario, the idea put forward is that mapping and merging of resources yields a practicable and viable solution to boost the effective exploitation of the available material.

A specific framework is presented that makes it possible to create new language resources via unification and packaging of different existing sources. The work presented here is in line with the activities promoted by ELRA towards the development of methods for packaging new language resources by combining independently created resources, and was carried out as part of the ELRA Production Committee (PCom) activities.

The focus of the paper is on the implementation of a specific procedure for merging monolingual lexicons: investigation is made on how one lexicon is obtained out of two, through a set of semi-automatic techniques that, on the basis of predefined mapping rules, first, detect candidate equivalent entries, next, merge them in a new unified entry. The immediate benefit consists of the integration of different types of information coming from external sources in a single resource.

The number of papers on unification of resources, be it lexicons or ontologies (Crouch & King, 2005; Ehrig & Sure, 2004), shows that this is an almost consolidated approach. Despite this, it is also true that no "merging protocol" for resource building exists and such a protocol is difficult to define for heterogeneous resources. The major contribution of the present work resides not solely in the implementation of the unification procedure per se,

but also in the proposal of an effective lexicon production model, which is attractive for producers and users: it could be called "unified lexica on demand", enabling the community to customize available individual language resources via unification.

The paper incorporates sound practices and results of previous experiments carried out at ILC-CNR (Monachini *et al.*, 2004) for producing new resources by combining two independently created ones. We describe the efforts to create a Unified Lexicon (UL) from two Italian computational lexicons, namely a common subset of lexical entries extracted from (i) the phonological layer of the PAROLE-SIMPLE-CLIPS lexicon (Ruimy *et al.*, 2002; Monachini *et al.* 2004)[1] and (ii) the LCSTAR pronunciation lexicon (Hartikainen *et al.*, 2003; Maltese & Montecchio, 2004)[2]. The objective of the merging procedure is to offer a newly packaged resource consisting of Unified Lexical Entries (ULEs), where heterogeneous linguistic information coming from source entries is contrasted, integrated, and, what's more important, via PAROLE, LCSTAR gains the access key to *hook* further layers of linguistic description (Mammini, Ulivieri & Monachini, 2005).

By exploiting a detailed mechanism for mapping grammatical classifications of UL entries, a set of Unified Morphosyntactic Specifications (UMS) shared by lexica for written and spoken Italian has been obtained as a *by-product* of the overall unification procedure. This constitutes a quite new return in this kind of activities, since (to our knowledge) it is the first time that morphosyntactic specifications for written and spoken lexica are compared and a consensual set is tried to reach. UMS represent a valuable contribution to activities aiming at defining standardized data categories for lexicons.

Merging practice is likely to have significant side-effects on cross-validation activities: the linguist, by analysing the entries that escape the mapping and are not candidate to the UL, the so-called *conflicts*, can detect not only mismatches due to diverging classifications in encoding but also probable errors in both resources.

2. Setting the conditions: pre-integration

When data from different sources need to be merged[3], the conversion of their native data structures and formats to a uniform structure and format is, unfortunately, only part of the battle, but, beyond doubt, one of the most time-consuming activities. Another crucial task is the identification of those parameters that detect equivalence between lexical entries in Lexicon^A and Lexicon^B and perform one-to-one mappings. All these stages constitute the so-called pre-integration**.** Finally, the mappings will be the input for the last step, where source entries candidate to the merging will be fused into one unified entry.

[1] http://www.ilc.cnr.it; see National Projects CLIPS.

[2] Project IST-2001-32216 LC-Star: Lexica and Corpora for Speech-to-Speech Translation Technologies, http://www.lc-star.com

[3] Due to the wide range of terms used in this area, we want to fix the distinction between *mapping* and *merging* and specify what we mean by those. *Mapping* identifies a relation such that an entry of Lexicon^A is associated with at least one entry of Lexicon^B. *Merging* is intended as combining together, so that the constituent parts of entry^a and entry^b form a new whole.

2.1. The test lexicon

In the case at hand, the experiment has been carried out on a test lexicon. In order to obtain this bench test, a sample of lexical entries common to both resources has been extracted. This sample has been composed with the following criteria in mind: (i) since we are dealing with pronunciation lexicons, homographic but not homophonous words, i.e. *fòrmica* vs. *formìca* (*Formica* vs. *ant*); *pèsca* vs. *pésca* (*peach* vs. *fishing*) have been considered a challenging test bed; (ii) since morphosyntactic information is part of the game, we aimed to have linguistic labels significantly represented; (iii) finally, half of the sample has been extracted taking into account frequency information obtained from the IT-PAROLE Corpus. In this way, we defined a common subset of 936 different orthographic forms, corresponding to 2226 LCSTAR entries (<entry> elements, differing for grammatical interpretation or phonetic transcription) and 2429 PAROLE entries (*Phonological Units*). Figure 1. shows the distribution of the starting subset of core lexical entries over different categories.

Pos	LCSTAR	PAROLE
Auxiliary	35	
Adjective	280	370
Adverb	59	66
Article	11	8
Conjunction	13	14
Determiner	52	58
Interjection		8
Noun	945	887
Numeral	14	37
Preposition	29	25
Pronoun	70	87
Verb	711	869[4]
Total	**2226**	**2429**

Figure 1. Distribution of test entries over categories

2.2. The two computational lexicons

2.2.1. PAROLE lexical entries

In PAROLE, the phonological layer is organized around the notion of phonological unit (PhU), the basic unit of this representational layer. The PhU is the entry-key of a word-form lexical entry. Along this layer, the splitting between different units is done according to the different linguistic features relevant at this descriptive level, i.e. pronunciation information: due to the two different phonological behaviors of, e.g., *pèsca* vs. *pésca*, two separate lexical entries exist for *pesca* (with two different identifiers: PhUpesca and PhUpesca2). Each PhU points to one or more basic unit(s) of the further layer, the morphological unit (MU) which provides the lemma, its morpho-syntactic features and the inflectional code. A PhU is represented in XML format; a Document Type Definition (DTD) defines its structure and describes content information.

[4] Verbs in PAROLE also include Auxiliary class.

2.2.2. LCSTAR lexical entries

In LCSTAR, the entry-point for accessing phonological and morphological information of lexical entries is based on the concept of <entrygroup>, which groups together all word-forms sharing the same spelling. The entrygroup refers to a generic spelling (word-form). For each entrygroup, the following is specified: <orthography>, i.e. the spelling, zero or more alternative spellings, one or more <entry> elements. For each entry, the following is specified: <PoS>, <lemma> and <phonetic> transcription. Concretely, *pesca* has a unique entrygroup, containing as many <entry> elements as possible grammatical interpretations linked to that orthography exist, with the relevant lemma and relevant phonetic transcriptions. Similarly to PAROLE, LCSTAR lexical entries are represented in XML format and a DTD (that covers all languages in the project) is provided.

2.2.3. Interoperability of the two formats

The first task of pre-integration stage, i.e. the selection of a uniform format in order to make the two sources formats interoperable, has been made easier by the fact that information included in both lexicons is coded with an XML-based mark-up language. It represents the linguistic information is a formal and unambiguous way and is both easy to read and to process. Once the data-model is defined, the same specialized automatic routines parsed the sample XML entries and imported them in a relational database.

2.3. Representation of linguistic content

In order to set an ideal mapping environment, the second obstacle to overcome is to make the representation formats of linguistic information interoperable as well. In this task, the adherence to standards for the representation of linguistic content is another variable which plays a crucial role.

2.3.1. Pronunciation

As far as pronunciation information is concerned, both lexicons provide overlapping information: position of stress, quantity of vowels and quality of consonants. Syllable boundaries are marked as well.

Both lexicons represent this information by adopting the SAMPA standard phonetic alphabet (www.phon. ucl.ac.uk/home/sampa/index.html). The total convergence of format and content at this layer represents a benefit in view of the interoperability of pronunciation information.

2.3.2. Morpho-syntactic information

At this level, both lexicons provide the same bulk of information: lemma, grammatical classification and morphological features. With respect to LCSTAR, PAROLE phonological entries, via the link to morphological unit(s) allow to access information about the inflectional paradigm (with the rules for generating the whole inflection) and the link to further layers of linguistic description.

As concerns the content, both word classification systems are compliant to the morpho-syntactic specifications issued by EAGLES (Monachini & Calzolari, 1996): this makes them highly compatible and generates high expectations about their successful mapping. The major outstanding difference is the surface representation of the two morphological encoding systems: PAROLE presents a set of labels, ascii strings built on the basis of the "character-position" strategy (Marinelli *et al.*, 2003), whereas LCSTAR prefers an attribute-value feature-structure notation expressed in XML[5].

2.3.3. Intermediate Format

The best way to handle these presentational differences, thus making the two representations directly matchable, was to adopt a strategy well tested in other similar initiatives and to translate the two formats in an internal notation (see Figure 2.). This language-neutral internal representation consists of the *Intermediate Format* (IF), defined for the first time in EAGLES (Leech & Wilson, 1999): it has been refined and used in a number of projects as a simple and easy way to allow different physical labels to map each other. The basic structure of this internal representation format is very straightforward and particularly fits for the purpose of automatic mapping. It is not intended to be used by human users, but only serve as an interchange mechanism. It consists of character-strings, where a character in specific position represents an attribute-value pair. Position 1 encodes part-of-speech (e.g. N = noun), position 2 encodes sub-category (Type: e.g. c = common), whereas position 3 and 4 encode the values relevant for morphological attributes (Gender: e.g. m = masculine; Number: e.g. s = singular)[6].

Lexicon	Morpho-syntactic label	IF
PAROLE	NMS	Ncms-
LCSTAR	PoS="NOM" type="common" gend="masculine" numb="singular"	Ncms-

Figure 2. IF for equivalent morpho-syntactic labels

The IF has been made more powerful in its interchange function: category by category, special mapping rules have been defined in order to allow two IFs with diverging surface strings to map nevertheless. These rules offer the advantage of maximizing the mapping of equivalent lexical entries with not exactly equivalent linguistic classification. Moreover, the indication of the degree of (dis)similarity has been provided by means of a *score*. Figure 3, 4 and 5, below, give the idea of different scores assigned to corresponding IFs, depending on the degree of their equivalence.

CG	PAR	LCS	*score*
NOUN	Ncfp-	Ncfp-	10
NOUN	Ncfs-	Ncfs-	10
NOUN	Ncmp-	Ncmp-	10
NOUN	Ncms-	Ncms-	10

Figure 3. Perfectly overlapping IFs

[5] It should be noted that this is only a pure physical notational difference, being the two systems perfectly translatable one into each other.

[6] This internal notation format is perfectly translatable in a feature structure system.

CG	PAR	LCS	*score*
NOUN	Ncfs-	Ncf_-	20
NOUN	Ncfp-	Ncf_-	20
NOUN	Ncmp-	Ncm_-	20
NOUN	Ncms-	Ncm_-	20

Figure 4. Not-overlapping IFs (L^A more specific)

CG	PAR	LCS	*score*
NOUN	Npfp-	N2f_-	40
NOUN	Npfp-	N3f_-	40
NOUN	Npfp-	N4f_-	40

Figure 5. Not-overlapping IFs (L^A and L^B complementary)

The conversion of formats from native formats to the IF is a very delicate task, due to time and manual efforts: it is far from being trivial, even for someone familiar with the resources, since it implies, not only deep knowledge of native data representation formats and of linguistic content encoded but also, and above all, of IF internal functioning. The definition of mapping rules between the two systems implies a careful manual control as well.

This phase plays a crucial role in the next step, i.e. the identification of Unified Morphosyntactic Specifications, which can be supplied with the indication of the degree of their correspondence (see section 3.3).

2.4. Mapping

The mapping consists of an automatic routine that, given mapping rules, compares two entries from $Lexicon^A$ and $Lexicon^B$ ($entry^a$ and $entry^b$) and tests their equivalence over a mapping window (Figure 5.).

MAPPING WINDOW			
orthography	*lemma*	*transcription*	*IF*

Figure 5. Mapping window

$Entry^a$ and $entry^b$ are considered equivalent and candidates to become an $entry^{UL}$, if all fields of the mapping window perfectly coincide.

In case the fields of IFs differ, the predefined mapping rules come into play, match IFs and allow two entries with same orthography, lemma, transcription but diverging IFs to map nevertheless. These rules also assign the *score* of (dis)similarity (see Figure 6.).

score	Type of (dis)similarity
10	L^A and L^B perfectly overlapping
11	overlapping (*invariant* in L^B)
20	L^A more specific
30	L^B more specific
31	L^B more informative (*invariant* in L^B)
100	(Sub-)category present in L^A
200	(Sub-)category present in L^B
40	L^A and L^B complementary info

Figure 6. Mapping scores

3. The Merging

At this point, the merging routine merges two candidate equivalent lexical entries into one Unified Lexical Entry (ULE). An ad-hoc defined Document Type Definition (DTD), a formally specified grammar, describes content information and defines the structure of the ULE (see Figure 7.) consisting of the following XML elements:

- a first bundle of unified information, orthography, lemma and phonetic transcription (the mapping parameters);
- identifiers of PAROLE phonological and morphological units, which allow LCSTAR entries to extract, if needed, inflectional codes for lemmas and, moreover, to hook further layers of linguistic representation, in particular the semantic descriptive level;
- finally, the bundle of mapped morphosyntactic specifications inherited from source entries, with the indication of their (dis)similarity.

```
<entry orthography="pesca" lemma="pesca" phonetic="' p e - s k a">
   <linkILCLexicon Phu="PHUpesca" Mus="MUSpescaNOUN2"/>
   <MappingFeature>
      <FeatureILC PoS="NOM" type="common" gender="feminine" number="sing
      <FeatureIBM PoS="NOM" class="common" gender="feminine" />
      <relation description="descr. 20"/>
   </MappingFeature>
</entry>
<entry orthography="pesca" lemma="pesca" phonetic="' p E - s k a">
   <linkILCLexicon Phu="PHUpesca2" Mus="MUSpescaNOUN"/>
   <MappingFeature>
      <FeatureILC PoS="NOM" type="common" gender="feminine" number="sing
      <FeatureIBM PoS="NOM" class="common" gender="feminine" number="sin
      <relation description="descr. 10"/>
   </MappingFeature>
</entry>
<entry orthography="pesca" lemma="pescare" phonetic="' p e - s k a">
   <linkILCLexicon Phu="PHUpesca" Mus="MUSpescareVERB"/>
   <MappingFeature>
      <FeatureILC PoS="VER" mode="indicative" number="singular" person="3" t
      <FeatureIBM PoS="VER" mood="indicative" number="singular" person="3"
      <relation description="descr. 40"/>
   </MappingFeature>
</entry>
<entry orthography="pesca" lemma="pescare" phonetic="' p e - s k a">
   <linkILCLexicon Phu="PHUpesca" Mus="MUSpescareVERB"/>
   <MappingFeature>
      <FeatureILC PoS="VER" mode="imperative" number="singular" person="2"
      <FeatureIBM PoS="VER" mood="imperative" number="singular" person="2"
```

Figure 7. An XML ULE

The merging consists of cyclical steps, with alternate phases of automatic mapping and human detection of conflicts, with the aim of increasing the coverage of UL, by correcting mismatches.

3.1. Unified Lexicon

With as input 2226 and 2429 source entries from LCSTAR and PAROLE, respectively, 2003 successful one-to-one-mappings verifies, which give rise to as many UL entries exportable in XML format.

ULEs are distributed over different PoS as appears in Figure 8. The figure also reports, category by category, the percentage of entries merged into the UL from both LCSTAR and PAROLE. For example, 306 UL adjective entries represent the 86,69% of total LCSTAR adjective entries and the 82,70% of total PAROLE entries.

Results seem very promising and generate high expectations in view of extending the unification procedure to a wider portion extracted from the two lexicons.

Pos	UL	LCSTAR	Parole
Adjective	306	86,69%	82,70%
Adverb	58	98,31%	87,88%
Article	7	63,64%	87,50%
Conjunction	13	100%	92,86%
Determiner	56	90,32%	96,55%
Interjection	7	100%	87,50%
Noun	796	62,27%	86,70%
Numeral	20	95,24%	54,05%
Preposition	25	46,30%	100%
Pronoun	79	96,34%	90,80%
Verb	663	85,66%	76,29%
Total	**2003**		

Figure 8. Distribution of UL entries over PoS

3.2. Unified Morphosyntactic Specifications

From the UL, a set of 205 Unified Morphosyntactic Specifications (UMS) for lexicons of written and spoken Italian can be exported. This set of XML feature-structures is an obvious and uncontroversial benefit for prospective users. The fact that equivalent/corresponding specifications from source lexicons appear juxtaposed in their native formats, carrying indication of their (dis)similarity is together a novelty and an added-value in this kind of mapping exercises.

This outcome can be seen as a valuable new contribution to activities undertaken within ISO-TC37/SC4, aiming at revising the ISO Registry of standardized data categories for lexicons (ISO 12620).

```
<MappingFeature>
   <FeatureILC PoS="NOM" class="common" gender="feminine" number="singular" />
   <FeatureIBM PoS="NOM" class="common" gender="feminine" number="singular" />
   <relation description="descr. 10"/>
</MappingFeature>
<MappingFeature>
   <FeatureILC PoS="NOM" class="common" gender="masculine" number="singular"/>
   <FeatureIBM PoS="NOM" class="common" gender="masculine" number="invariant" />
   <relation description="descr. 11"/>
</MappingFeature>
<MappingFeature>
   <FeatureILC PoS="NOM" class="common" gender="masculine" number="singular"/>
   <FeatureIBM PoS="NOM" class="common" gender="masculine" number="singular" />
   <relation description="descr. 10"/>
</MappingFeature>
<MappingFeature>
   <FeatureILC PoS="NOM" type="common" gender="feminine" number="plural" />
   <FeatureIBM PoS="NOM" class="common" gender="feminine" number="invariant" />
   <relation description="descr. 11"/>
</MappingFeature>
<MappingFeature>
   <FeatureILC PoS="NOM" type="common" gender="feminine" number="plural" />
   <FeatureIBM PoS="NOM" class="common" gender="feminine" number="plural" />
   <relation description="descr. 10"/>
</MappingFeature>
<MappingFeature>
   <FeatureILC PoS="NOM" type="common" gender="feminine" number="plural" />
   <FeatureIBM PoS="NOM" class="common" gender="invariant" number="plural"/>
   <relation description="descr. 11"/>
```

Figure 9. Examples of XML UMS.

A further extension of this experiment could be to test how merging results vary changing the mapping window: by mapping "all fields but the lemma", different lemmatization policies can be detected; vice-versa, by mapping "all fields except for phonetic" different phonetic transcriptions can emerge. This could be also a good strategy to increase the coverage of UL.

3.3. *Conflicts*

Conflicts represent cases of mismatches in the mapping. They are a very interesting outcome for the linguist, since they can point out inconsistencies in the encoding strategy of the two lexicons, in particular, diverging attribution of (i) PoS to the same lemma (*difficile* as noun in LCSTAR and as adjective in PAROLE), (ii) lemma to the same word-form (*leonessa, gatta* under the lemmas *leone* and *gatto*, respectively, in LCSTAR, but under *leonessa* and *gatta* in PAROLE), (iii) subcategory to the same lemma and PoS (*nord* as proper-noun in PAROLE and common-noun in LCSTAR).

From the analysis of conflicts, possible errors in word classification or cases of missing encoding in either resource can also emerge (e.g. *fa* encoded as noun – the musical note – was missing in PAROLE). This has strong impacts on the value of merging in cross-validation of resources.

4. A Production paradigm: Lexica on demand

Experiments in monolingual lexicon merging taught that one the most time-consuming task is pre-integration. Standards for encoding linguistic content constitute a crucial variable that can ease conversion/translation steps. Based on procedures adopted, we can conclude with suggestions that could make this production process attractive to resource producers.

As an outcome of the UL, ELRA has now the necessary expertise to provide, through its Production Network, appropriate services as a response to the need for merged and/or combined lexica. The idea is to offer, in addition to the catalogued resources, lexica that could be obtained through unification of several catalogued ones or even through the unification of partner's lexica with others from the ELRA catalogue. In addition to the technical aspects that will have to be tackled as proposed in this project, the legal aspects will also be properly addressed. This would boost the packaging or customization of existing lexical resources, as indicated by current needs in the language engineering community.

5. Conclusion: a new ELRA service

The ILC team and PCom have investigated the possibility of unified morphosyntactic specifications for lexica of written and spoken Italian (starting from a common sample of two Italian lexicons: PAROLE/SIMPLE/CLIPS and LCSTAR).

At this point, the merging routine has unified two equivalent/corresponding entries into one Unified Lexical Entry (ULE), thus producing the Unified Lexicon (UL). The merging consisted of cyclical steps, with alternate phases of automatic mapping and manual detection of conflicts, in the aim of augmenting the coverage of UL, by correcting mismatches. Once the merging results have been considered stable, DTDs have been defined and the relevant procedure implemented, in order to export the UL and the unified set of morphosyntactic labels in common to the two lexicons in XML format.

ELRA is now in a position to promote and support the development of methods for packaging new language resources by combining independently created resources.

ELRA aims to exploit the UL experience to carry out such mapping/merging activities for resources available on the ELRA catalogue, as well as those to be supplied by customers with the goal to provide unified on-demand lexicon databases. Such services will increase the flexibility of the ELRA offerings as well as the benefits of the ELRA customers providing them with better services around lexicon databases. In addition they will capitalize on the different networks set up by ELRA and its Production Committee (PCom).

6. References

Crouch, D. and King T.H. (2005). Unifying Lexical Resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany.

Ehrig, M. and Sure Y. (2004). Ontology mapping - an integrated approach. In *Proceedings of the First European Semantic Web Symposium, ESWS 2004*, volume 3053 of Lecture Notes in Computer Science, Heraklion, Greece. Springer Verlag, pp. 76-91.

Hartikainen, E., Maltese G., Moreno A., Shammass S., Ziegenhein U. (2003). Large Lexica for Speech-to-Speech Translation: From Specification to Creation, *Eurospeech 2003*, Geneva, Switzerland, September 2-7 2003, vol. 1, pp. 21-24.

ISO 12620. Terminology and other language resources – Data Categories – Data Category selection (DCS) for electronic lexical resources (ELR). ISO TC37/SC4.

Leech, G.N. and Wilson A. (1999). Standards for Tagsets. In *van Halteren H. (Ed.), Syntactic Wordclass Tagging*, pp. 55-80. Kluwer, Dordrecht.

Maltese, G. and Montecchio C. (2004). General and Language-specific specifications of contents of lexica in 13 languages. LCSTAR Deliverable D2.1, D2.2, D2.3, D2.4, v.2.1, IBM, Italy, pp. 60.

Mammini, M., Ulivieri M., Monachini M. (2005). Unified Lexica: Common sample lexicon and harmonized morpho-syntactic specifications between PAROLE and LCSTAR. Unified Lexica Deliverable (PCom ELRA), CNR-ILC, Pisa.

Marinelli, R., Biagini L., Bindi R., Goggi S., Monachini M., Orsolini P., Picchi E., Rossi S., Calzolari N., Zampolli A. (2003). The Italian PAROLE corpus: an overview. In *A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale*, Special Issue, XVI-XVII. Pisa-Roma, IEPI. Tomo I, pp. 401-421.

Monachini, M., Calzolari N. (1996). Synopsis and Comparison of Morphosyntactic Phenomena encoded in Lexicons and Corpora and Applications to European Languages, EAGLES Recommendations, Pisa (www.ilc.cnr.it).

Monachini, M., Calzolari F., Mammini M., Rossi S., Ulivieri M. (2004). Unifying Lexicons in view of a Phonological and Morphological Lexical DB. In *Lino M.T. et al. (eds.) Proceedings of LREC2004 – IV International Conference on Language Resources and Evaluation*, Lisbon, Portugal, vol. 3, pp.1107-1110.

Ruimy, N., Monachini M., Distante R., Guazzini E., Molino S., Ulivieri M., Calzolari N., Zampolli A. (2002). CLIPS, A Multil-level Italian Computational Lexicon: a Glimpse to Data. In *Proceeding of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas de Gran Canaria, Spain, vol III, pp.792-799.