Hybrid Method for Stress Prediction Applied to GLAFF-IT, a Large-Scale Italian Lexicon

Basilio Calderone, Matteo Pascoli, Franck Sajous, Nabil Hathout

Hybrid method for stress prediction applied to GLAFF-IT, a large-scale Italian lexicon

Basilio Calderone, Matteo Pascoli, Franck Sajous, and Nabil Hathout

CLLE-ERSS (CNRS & Université de Toulouse 2),
Maison de la Recherche - 5, allées Antonio Machado
F - 31058 Toulouse Cedex 9
name.surname@univ-tlse2.fr

Abstract. This paper presents a hybrid method for automatic stress prediction that we apply to GLAFF-IT, a large-scale Italian lexicon we extracted from GLAW-IT, a Machine-Readable Dictionary grounded on Wikizionario. Our approach combines heuristic rules and a logistic model trained on the words' sets of phonological features. This model reaches a 98.1% accuracy. The resulting resource is a large lexicon for the Italian language that we release under a free licence. It includes morphological and phonological information for each of its $457,702$ entries. As of today, it is the only Italian lexicon featuring both large coverage and indication of stress position.

Keywords: Italian stress prediction, phonological transcriptions, free large-scale lexicon, Wiktionary, Wikizionario

1 Introduction

As a consequence of the expansion of corpus and data-driven approaches to language, lexical resources (LRs) are nowadays essential for quantitative and qualitative studies of the language. Despite the linguistic richness available in existing LRs (morphological, morpho-syntactic annotations, semantic relations, etc.), phonological information such as the phonemic transcriptions of lexical forms and stress markers, is often not reported by such resources. This lack is problematic for data-based phonological analysis of the language and for phonetic and prosodic studies. Phonologists are interested in the sounds that are distinctive in a given language and on the rules that govern these sounds. In this context, the availability of a phonological lexicon tagged with stress placement is a prior condition for any investigations in the phonological domain. Phonological and stress information are also necessary in psycholinguistics where researchers manipulate a large set of word properties in order to design experimental protocols. Moreover, phonological lexicons reporting word stress and prosodic information are crucial in language acquisition analysis [12] and in the study of word recognition [16]. In a more practical perspective, phonological lexicons may integrate with other modules in NLP applications as, for example, text-to-speech systems [7].

Most resources conceived for the Italian language do not provide any word stress information and, more generally, they do not report phonological transcriptions at all [17, 4, 3, 14]. An exception is represented by PhonItalia [8], an Italian lexicon designed

for researchers working in the psycholinguistic domain. Besides the orthographic forms and their lemmas, PhonItalia also reports the phonological encoding of words with the stress placement. Although the resource provides a comprehensive range of phonological and distributional information, its limited coverage ($120,000$ entries) constitutes a serious deterrent for its utilisation in quantitative/descriptive language analysis and for its exploitation in the NLP domain.

In this paper, we present a hybrid method for the prediction of Italian stress and we apply it to a large-scale morpho-phonological lexicon. Our method combines phonologically-motivated rules with a logistic regression model (henceforth *logit* model) for the automatic prediction of stressed/unstressed vowels. By exploiting this method we enrich the phonological transcriptions present in a large Italian lexicon, GLAFF-IT, with the word stress placement. Besides the lexical resource itself, a significant contribution of this work is the method we developed for stress prediction. Used here to complete the current version of the lexicon, the method will also prove useful in the future to generate the transcriptions and stress placements of the neologisms that are regularly added to Wikizionario (or potentially originating from other sources). The paper describes the creation of GLAFF-IT from a machine-readable dictionary that encodes the Wikizionario's micro- and macrostructure (section 2). We present the lexicon and explain how we complete it with the missing forms by exploiting the systematic regularities of the Italian language regarding the orthography and the phonology. Section 3 focuses on the phonological transcriptions of GLAFF-IT and the problematic issue of Italian stress. In this section, we propose a hybrid method adopted for the stress assignment. The evaluation of the predictions and some conclusions are given in section 4.

2 GLAFF-IT, a large-scale Italian lexicon

2.1 GLAW-IT

In a previous paper [6], we introduced GLAW-IT, a free machine-readable dictionary (MRD) that encodes in a workable XML format the micro- and macrostructure of Wikizionario, the Italian edition of Wiktionary.[1] The method used to convert Wikizionario into GLAW-IT is similar to the conversion process from Wiktionnaire (the French language edition of Wiktionary) to GLAWI [15, 10]. GLAW-IT contains all the lexical knowledge found in Wikizionario. Articles may include fields reporting etymologies, definitions, lemmas and inflected forms, lexical semantic and morphological relations, hyphenations, translations and phonological transcriptions. GLAW-IT does not report systematically all this kind of information. The basic unit of GLAW-IT is the wordform and when some homographs correspond to the same wordform, the article contains a separate POS section for each one of them. An example is reported in Figure 1 for the entry *danno*, which is both the lemma (masculine singular) of *danno* 'damage' and the inflected forms (indicative present, 3rd person plural) of the verb *dare* 'to give'. As we can see, the stress placement is reported independently in the hyphenation field (*dàn|no*) as well as in the phonological transcription (ˈdaːnno). It may also occur in only one of the two fields or be missing. Conceived as a general-purpose MRD, GLAW-IT

[1] Available at: `http://redac.univ-tlse2.fr/lexicons/glawit.html`

```
<article>
  <title>danno</title>
  <id>70444</id>
  <text>
    <etymology>
      <wiki>dal [[latino]] ''[[damnum]]''</wiki>
      <xml>dal <innerLink>latino</innerLink> <i><innerLink>damnum</innerLink></i></xml>
      <txt>dal latino damnum</txt>
    </etymology>
    <pos type="sost" lemma="1" m="1" sing="1">
      <paradigm>
        <inflection type="Ncmp" form="danni"/>
        <inflection type="Ncms" form="danno"/>
      </paradigm>
      <defs>
        <def>
          <gloss>
            <labels>
              <label type="term">diritto</label>
              <label type="term">economia</label>
            </labels>
            <wiki>{{term|diritto|it}}{{term|economia|it}} conseguenza di un'[[azione]]
                negativa subita</wiki>
            <txt>conseguenza di un'azione negativa subita</txt>
            <xml>conseguenza di un'<innerLink>azione</innerLink> negativa subita</xml>
          </gloss>
        </def>
        <def>
          <gloss>
            <labels> <label type="term">medicina</label> </labels>
            <wiki>{{Term|medicina|it}} ogni fenomeno patologico che modifichi l'organismo o l'[[efficienza]]
              di una parte del corpo</wiki>
            <txt>ogni fenomeno patologico che modifichi l'organismo o l'efficienza di una parte del corpo</txt>
            <xml>ogni fenomeno patologico che modifichi l'organismo o l'<innerLink>efficienza</innerLink>
              di una parte del corpo</xml>
          </gloss>
        </def>
      </defs>
    </pos>
    <pos type="verb" lemma="0">
      <defs>
        <def>
          <gloss>
            <wiki>terza persona plurale, indicativo presente di [[dare]]</wiki>
            <txt>terza persona plurale, indicativo presente di dare</txt>
            <xml>terza persona plurale, indicativo presente di <innerLink>dare</innerLink></xml>
          </gloss>
        </def>
      </defs>
      <inflectionInfos>
        <inflectedForm gracePOS="Vmip3p-" lemma="dare"/>
      </inflectionInfos>
    </pos>
    <section type="sill"> <item>dàn|no</item> </section>
    <section type="pron"> <item type="IPA">'da:nno</item> </section>
    <section type="sin">
      <item>danneggiamento</item>
      <item>rottura</item>
      <item>oltraggio</item>
      <item labels="medicina">male</item>
      <item labels="medicina">lesione</item>
      <item>svantaggio</item>
    </section>
    <section type="ant">
      <item>sistemazione</item>
      <item>risarcimento</item>
    </section>
    <section type="der">
      <item>dannare</item>
    </section>
  </text>
</article>
```

Fig. 1. General structure of the article *danno* in GLAW-IT

is intended to be to used as such or as a starting point to tailor specific lexicons. In the section below, we explain how we derived GLAFF-IT (which stands for *Un Grande Lessico 'Tuttofare' dell'Italiano*, 'A Large Versatile Italian Lexicon') from GLAW-IT.

2.2 From GLAW-IT to GLAFF-IT

A first step in the creation of GLAFF-IT is filtering GLAW-IT by all the morphosyntactic and phonological tags and collecting all the inflected forms related to each lemma. In GLAW-IT, for a given lemma not all the inflected forms of the paradigm are present. This compels us to complete the missing forms by exploiting the systematic variations of Italian inflection with respect to each grammatical class. In particular, given some orthographic preconditions, certain inflected forms are totally predictable from other forms of the same paradigm. For example, for a particular set of missing nouns, we have generated the Plural Masculine forms from their Singular forms. Specifically, we have applied the general rule governing the alternation Masculine Singular/Masculine Plural for those nouns ending in *-o*, like *allomorfo* ('allomorph'), which change the last letter in *-i* for the Plural (Masculine) forms (*allomorfi*, 'allomorphs'). We report below the main deterministic inflection rules we implemented for the generation of the missing forms:

Nouns:
- lemma's Feminine Singular ending = *-a* → *-e* for the Feminine Plural, as in *casa* 'home' → *case* 'homes'
- lemma's Feminine Singular ending = *-e* → *-i* for the Feminine Plural, as in *siepe* 'hedge' → *siepi* 'hedges'
- lemma's Feminine Singular ending = *-ista* → *-iste* for the Feminine Plural, as in *rivista* 'magazine' → *riviste* 'magazines'

Adjectives:
- lemma's (Singular Masculine) ending = *-ico* → *-ici, -ica, -iche* respectively for the Masculine Plural, Feminine Singular and Plural, as in *magnifico* 'magnificent' → *magnifici*, *magnifica*, *magnifiche*
- lemma's (Singular Masculine) ending = *-go* or *-co* (but not *-ico*) → *-(c)(g)hi, -(c)(g)a, -(c)(g)he* respectively for the Masculine Plural, Feminine Singular and Plural, as in *metallico* 'metallic' → *metallici*, *metallica*, *metalliche*

Verbs:
- for the highly regular conjugation *-are*, we generate the missing verbal forms by adding regular inflectional suffixes to the stem base of the verb (which is always detectable). For example the Gerund and the Singular and Plural Present Participle of the verb *manipolare* 'to manipulate' are created by adding respectively *-ando*, *-ante* and *-anti* to the stem base *manipol*: *manipolando*, *manipolante* and *manipolanti*. When missing, we create 54 verbal forms of the paradigm by exploiting the regular inflectional suffixes of the Italian conjugation.

The aforementioned inflection rules enabled us to generate $42,845$ new wordforms which are integrated into GLAFF-IT. Table 1 reports the number of lemmas extracted from GLAW-IT and the number of forms generated by the aforementioned rules. The

	LEMMAS	FORMS		
		Initial	Generated	Total
Nouns	19,340	36,726	2,505	39,231
Adjectives	7,835	23,932	4,140	28,072
Verbs	7,552	351,604	36,200	387,804
Adverbs	2,593	2,595	0	2,595
Total	37,320	414,857	42,845	457,702

Table 1. Size of GLAFF-IT: number of lemmas and forms

current version of GLAFF-IT counts $37,320$ lemmas for $457,702$ wordforms and includes nouns, verbs, adjectives and adverbs.[2] Each entry of the lexicon includes a wordform, a tag in MULTEXT-GRACE format [13] specifying the main syntactic category and inflection features, a lemma and API phonological transcriptions with the stress placement when present in GLAW-IT. An extract of GLAFF-IT is reported in Figure 2. This version has been converted into Lexical Markup Framework, as illustrated in Figure 3.

```
...
danni|Ncmp|danno|ˈdaːnni
danno|Ncms|danno|ˈdanno
danno|Vmip3p-|dare|ˈdanno
dannoso|Afpms-|dannoso|dannˈoso
dannosa|Afpfs-|dannoso|dannˈosa
dannose|Afpfp-|dannoso|dannˈose
...
```

Fig. 2. An extract of GLAFF-IT.

As is the case for the inflected forms, phonological information may be absent from GLAW-IT. The next section describes the automatic generation of missing phonological transcriptions and stress placements.

3 The problematic Italian stress issue

3.1 Phonological transcriptions and stress in GLAW-IT

Only 4.2% of GLAW-IT's wordforms (corresponding to $17,720$ articles) include phonological transcriptions. 98.5% of these transcriptions also report stress placement. The stress information has been taken either from the phonological field or from the hyphenation field (cf. section 2.1). Table 2 provides a breakdown with respect to the four grammatical classes.

[2] GLAFF-IT is freely available at: `http://redac.univ-tlse2.fr/lexicons/glaffit.html`

```
<lexicalEntry id="danno">
  <formSet>
    <lemmatizedForm>
      <orthography>danno</orthography>
      <grammaticalCategory>commonNoun</grammaticalCategory>
      <grammaticalGender>masculine</grammaticalGender>
      <pronunciation>'da:nno</pronunciation>
    </lemmatizedForm>
    <inflectedForm>
      <orthography>danno</orthography>
      <grammaticalNumber>singular</grammaticalNumber>
      <pronunciation>'da:nno</pronunciation>
    </inflectedForm>
    <inflectedForm>
      <orthography>danni</orthography>
      <grammaticalNumber>plural</grammaticalNumber>
      <pronunciation>'da:nni</pronunciation>
    </inflectedForm>
  </formSet>
</lexicalEntry>
<lexicalEntry id="dannoso">
  <formSet>
    <lemmatizedForm>
      <orthography>dannoso</orthography>
      <grammaticalCategory>adjective</grammaticalCategory>
      <pronunciation>da:nno'so</pronunciation>
    </lemmatizedForm>
    <inflectedForm>
      <orthography>dannoso</orthography>
      <grammaticalGender>masculine</grammaticalGender>
      <grammaticalNumber>singular</grammaticalNumber>
      <pronunciation>da:nno'so</pronunciation>
    </inflectedForm>
    <inflectedForm>
      <orthography>dannosi</orthography>
      <grammaticalGender>masculine</grammaticalGender>
      <grammaticalNumber>plural</grammaticalNumber>
      <pronunciation>da:nno'si</pronunciation>
    </inflectedForm>
    <inflectedForm>
      <orthography>dannosa</orthography>
      <grammaticalGender>feminine</grammaticalGender>
      <grammaticalNumber>singular</grammaticalNumber>
      <pronunciation>da:nno'sa</pronunciation>
    </inflectedForm>
    <inflectedForm>
      <orthography>dannose</orthography>
      <grammaticalGender>feminine</grammaticalGender>
      <grammaticalNumber>plural</grammaticalNumber>
      <pronunciation>da:nno'se</pronunciation>
    </inflectedForm>
  </formSet>
</lexicalEntry>
```

Fig. 3. Extract of GLAFF-IT in Lexical Markup Framework

	Total	Stressed	Without stress
Nouns	9,135 (23.28%)	8,983 (22.89%)	152 (0.38%)
Adjectives	4,227 (15.05%)	4,196 (14.94%)	31 (0.11%)
Verbs	4,165 (1.07%)	4,093 (1.05%)	72 (0.01%)
Adverbs	193 (7.43%)	188 (7.24%)	5 (0.19%)
Total	17,720	17,460	260

Table 2. Number and percentage of phonological transcriptions with and without stress in GLAW-IT

In order to assess GLAW-IT's phonological transcriptions, we compare them to those of PhonItalia. We first build the intersection of their entries, resulting in $66,371$ orthographic forms (that corresponds to 15% of GLAW-IT and 57% of PhonItalia's vocabularies). The number of wordforms per POS contained in this intersection, as well as the proportion of forms having transcriptions and stress placements in GLAW-IT, are reported in Table 3.

	Wordforms	with phonological transcription in GLAW-IT		
		Total	with stress	without stress
Nouns	18,315	6,626 (36.18%)	6,535 (35.68%)	91 (0.5%)
Adjectives	12,835	2,763 (21.53%)	2,744 (21.4%)	19 (0.1%)
Verbs	34,233	1,867 (5.45%)	1,823 (5.33%)	44 (0.13%)
Adverbs	988	154 (15.59%)	149 (15.08%)	5 (0.5%)
Total	66,371	11,410	11,251	159

Table 3. Intersection of GLAW-IT and PhonItalia.

We then compared the 11,410 shared entries in order to check the correspondence of their transcriptions and stress placements. The results of the comparison are given in Table 4: GLAW-IT and PhonItalia present a 87% agreement both for the transcriptions and the stress placement. The main differences stem from the inventory of phonems used by the two resources. For example, in the group of the nasal consonants, GLAW-IT marks the velar nasal /ŋ/ (as in <anche> 'also' /aŋke/), which is absent in PhonItalia (where <anche> is transcribed /anke/). In such a case, our comparison, based on the exact matching of phonemes, leads to a disagreement.

3.2 Automatic generation of phonological transcriptions in GLAFF-IT

This section describes a method to generate phonological transcriptions from orthographic forms. We used this method to transcribe each wordform from GLAFF-IT when no transcription is given in GLAW-IT.

Knowing the pronunciation of an Italian word is enough to know its orthography (with very few exceptions, e.g. /kw/ in some words is written as <cu> - e.g. <cuore> 'heart' /kwɔre/ - instead of the more common <qu> - e.g. <quale> 'which' /kwale/).

	Agreement between GLAW-IT and PhonItalia			
	# entries	phonological transcription	# entries	stress placement
Nouns	6,626	5,635 (85.05%)	6,535	5,585 (85.46%)
Adjectives	2,763	2,448 (88.59%)	2,744	2,442 (88.99%)
Verbs	1,867	1,697 (90.89%)	1,823	1,669 (91.55%)
Adverbs	154	137 (88.96%)	149	132 (88.59%)
Total	11,410	9,917 (86.91%)	11,251	9,828 (87.35%)

Table 4. Agreement between GLAW-IT and PhonItalia with respect to the phonological transcriptions and the stress placement.

The opposite is not true in general: if most of the Italian graphemes are mapped one-to-one to phones, some cases do not allow automatic conversion. Here are some representative examples:

- a written <e> can be realised as /e/ or /ɛ/ when stressed. This distinction is crucial because it enables to differentiate homographic words having the same stress such as <pesca> = /pˈeska/ 'fishing' and /pˈɛska/ 'peach'
- a written <o> can be realised as /o/ (as in <asino> 'donkey' /azino/) or /ɔ/ when stressed (as in <rosa> 'rose' /rɔza/)
- a written <z> can be the voiced affricate /dz/ as in <zaino> 'backpack' /dzaino/ or unvoiced affricate /ts/ as in <canzone> 'song' /kantsone/
- a written <s>, intervocalic or after suffixes, can be realised as /s/ or /z/ as <casa> = /kasa/ 'home', <rosa> = /rɔza/ 'rose', <transatlantico> = /tranzˈatlantico/ 'transatlantic', <transiberiano> = /tranzˈsiberiano/ 'trans-Siberian')
- <j>, <y> and <w>, mostly found in loanwords (as in 'jazz', 'yacht', 'know-how'), can represent a wide range of phonemes: /j/, /dʒ/, /ʒ/, /i/, /v/, /w/, etc.

To set up a letter-to-phoneme mapping, henceforth *orth2phon*, we distinguished unambiguous from ambiguous cases:

Unambiguous cases We adopted the conversion rules grapheme(s)-to-phoneme only for the unambiguous cases in which we distinguished transparent and opaque conditions. For example, some cases of transparent grapheme-to-phoneme mapping are given by <p> → /p/ (<pane> 'bread' /pane/), <l> → /l/ (<lino> 'linen' /lino/) or <t> → /t/ (<tutto> 'all' /tutto/). Less transparent cases of mapping are <gli> → /ʎ/ (<figlio> 'son' /fiʎʎo/) and <gli> → /gl/ (<siglare> 'to initial' /siglare/). Conversely, opaque cases need the orthographic context to be converted in phonemes. Opaque cases are for example <ch> → /k/ (<anche> 'also' /aŋke/), <q> → /k/ (<quadri> 'paintings' /kwadri/), <c> → /k/ (<casa> 'home' /casa/). In many cases the stressed vowel reported in the hyphenation information can be used for the disambiguation of some ambivalent phonemes such as <e> or <o> that are respectively realised as /ɛ/ and /ɔ/ when stressed.

Ambiguous cases We encoded by a *ad hoc* capital letter all the cases which are not unambiguously convertible. For example, an intervocalic <s> may lead to the two different phonemes /s/ or /z/ and cannot be automatically predicted. We choose

to encode such cases into capital letters (here, intervocalic <s> → S). We define eight different ambiguous cases:
1) E = /e/ or /ɛ/; 2) O = /o/ or /ɔ/; 3) I = /i/, /j/ or ∅; 4) U = /u/ or /w/; 5) S = /s/ or /z/; 6) Z = /ts/ or /dz/; 7) J = /j/ or /dʒ/; 8) W = /w/ or /v/

We first applied *orth2phon* to the entries from the intersection between GLAW-IT and PhonItalia described in Table 3. We then compared the agreement between the two resources on their phonological transcriptions and stress placements. As can be seen in the Table 5, the *orth2phon* coding increases GLAW-IT and PhonItalia's agreement on transcriptions from 86% (cf. Table 4) to 92% and the agreement on stress placement (from 86% to from 90%).

	Agreement between GLAW-IT+*orth2phon* and PhonItalia			
	# entries	phonological transcription	# entries	stress placement
Nouns	6,626	6,056 (91.39%)	6,535	5,753 (88.03%)
Adjectives	2,763	2,581 (93.41%)	2,744	2,528 (92.12%)
Verbs	1,867	1,755 (94.00%)	1,823	1,669 (91.55%)
Adverbs	154	137 (88.96%)	149	132 (88.59%)
Total	11,410	10,529 (92.27%)	11,251	10,082 (89.60%)

Table 5. Agreement between GLAW-IT+*orth2phon* and PhonItalia intersection.

Finally, the *orth2phon* method enabled us to generate all the phonological transcriptions missing from GLAFF-IT. In the next section we introduce our method for the prediction of the stress placement in the phonological transcriptions of GLAFF-IT.

3.3 A hybrid method for stress prediction in GLAFF-IT

Stress placement is lexically marked in Italian [11, 1] and has a contrastive function: *capito* /ˈkapito/ 'I happen by' vs. *capito* /kaˈpito/ 'understood' vs. *capitò* /kapiˈto/ 'it happened'. This means that *a priori* the speaker should have a phonological knowledge of the word to pronounce it correctly (unless the stress is on the last vowel: in this case, the stress is orthographically marked as in *capitò*).

As Figure 1 shows, GLAW-IT can report word stress information in two different sections: a) in the phonological transcriptions and b) in the hyphenation of the word in which the stressed vowel is orthographically marked. By gathering information from both these sections we collect $59,891$ wordforms presenting a stress marker (Table 6). In this section, we present a hybrid method which allows us to complete the stress information for the remaining $397,812$ wordforms of GLAFF-IT by combining heuristic rules and predictions performed by a *logit* model.

Heuristic rules

In some known cases, stress placement in Italian is deterministic. We design a set of heuristic rules in order to predict words' stress in such situations:

- Stress position is generated for the words in which the stressed vowel is orthographically marked, as in <liquidità> 'liquidity' /likwiditˈa/ or <avrò> 'I will have' /avrˈo/.
- Bisyllabic words in which the last vowel is not graphically marked always have the stress on the first vowel, as <gonna> 'skirt' /gˈɔnna/ or <casa> 'home' /kˈasa/.
- Some verbs with particular endings have regular stress patterns. For example, it is the case of the verbal ending /-vˈamo/ which identifies the 1st person imperfect indicative as in <amavamo> 'we were loving' /amavˈamo/ or the ending of 3rd person present conditional /-rˈɛbbero/ as in <amerebbero> '(they) would love' /amerˈɛbbero/. We distinguished 15 verbal endings (mostly coming from the subjunctive and conditional mood) exhibiting predictable stressed vowels.

By exploiting the heuristic rules, we determine the stress placement for $289,717$ forms. Table 6 reports the details of this stress generation. This processing mainly involves the verbal entries ($288,095$ stress placements generated) and a limited number of nouns and adjectives. This fact is quite unsurprising, given that Italian stress is lexically marked and so the application of possible heuristics is highly constrained for nouns and adjectives.

	Initially stressed	**Heuristic-generated**	**Total**	**Remaining to stress**
Nouns	24,435 (62.28%)	1,072 (2.73%)	25,507 (65.01%)	13,724 (34.98%)
Adjectives	17,908 (63.79%)	507 (1.80%)	18,415 (65.59%)	9657 (34.40%)
Verbs	15,144 (3.90%)	288,095 (74.28%)	303,239 (78.19%)	84,565 (21.80%)
Adverbs	2,403 (92.60%)	43 (1.65%)	2446 (94.25%)	149 (5.74%)
Total	59,891 (13.80%)	289,717 (63.29%)	349,607 (76.38%)	108,095 (23.60%)

Table 6. Number and percentage of the stress markers present in GLAW-IT and generated by heuristic rules.

Machine learning

Stress prediction has been performed for the $108,095$ remaining wordforms for which no heuristic rule has been applied. We use a model that has learned the phonological contexts for stressed and unstressed Italian vowels. Orthographic and phonological context-based approaches have been extensively used in the text-to-speech domain for stress detection [2, 7] and for accenting unknown words in a specialised language [18]. The rationale behind our approach is that the exploitation of the phonological neighbourhood of a vowel helps estimate its probability of being stressed or unstressed.

We trained a feed-forward (one hidden layer) neural network, using a logistic activation function, encoding the phonological neighbourhoods of the unstressed and the stressed vowel (respectively 0 and 1). Such a method has been used successfully in similar settings for the syllabification of Italian words [5]. In our model, the representation of the input data was constituted by each vowel composing the word with its left and right phonological context. The Figure 4 reports the representation for two Italian words, <decadere> 'to decay' /dekadˈere/ and <decidere> 'to decide' /detʃˈidere/. Although

ID	L_n	$L_{...}$	L_8	L_7	L_6	L_5	L_4	L_3	L_2	L_1	**Focus**	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	$R_{...}$	R_n	**Out**
1	#	#	#	#	#	#	#	#	#	d	e	k	a	d	e	r	e	#	#	#	#	→ 0
2	#	#	#	#	#	#	#	d	e	k	a	d	e	r	e	#	#	#	#	#	#	→ 0
3	#	#	#	#	#	d	e	k	a	d	e	r	e	#	#	#	#	#	#	#	#	→ 1
4	#	#	#	d	e	k	a	d	e	r	e	#	#	#	#	#	#	#	#	#	#	→ 0
1	#	#	#	#	#	#	#	#	#	d	e	tʃ	i	d	e	r	e	#	#	#	#	→ 0
2	#	#	#	#	#	#	#	d	e	tʃ	i	d	e	r	e	#	#	#	#	#	#	→ 1
3	#	#	#	#	#	d	e	tʃ	i	d	e	r	e	#	#	#	#	#	#	#	#	→ 0
4	#	#	#	d	e	tʃ	i	d	e	r	e	#	#	#	#	#	#	#	#	#	#	→ 0

Fig. 4. Input representation for two Italian words, <decadere> (/dekadˈere/) 'to decay' and <decidere> (/detʃˈidere/) 'to decide' .

phonologically quite similar and with identical syllabic structure, the two words present different stress placement: on the third vowel for <decadere> and on the second for <decidere>. In the input representation, the binary response unstressed/stressed vowel (0/1) is mapped to the left (L) and right (R) phonological context of the vowels (which are in Focus position). Each context defines the phonemes occurring in a given position with respect to the vowel in Focus. For example, the L_1 and R_1 contexts indicate respectively the phonemes occupying the first position on the left and the first position on the right with reference to the Focus. There are as many rows as the number of vowels in the word. The number of contexts considered is determined by the phonological length of the words: the longest word imposes the final number of contexts that will be equal to its number of phonemes N - 1 for L and R. In our input representation, each phoneme is encoded as a set of binary features defining the place and the manner of articulation. Although some authors report a clear correlation between stress patterns and phonological similar words [9], our choice was largely motivated by phonological reasons, with the rationale of taking into account the specific phonemic nature of each phoneme according to a set of phonologically-based features. We distinguished 14 features for all the Italian phonemes:

1. Voicing (VO): marks that the phonemes is voiced - or not
2. Bilabial (BI): consonants articulated with both lips
3. Labiodentals (LD): consonants articulated with the lower lip and the upper teeth
4. Dental-alveolar (DA):consonants articulated with a flat tongue against the alveolar ridge and upper teeth
5. Palato-alveolar (PA): consonants articulated with the blade of the tongue behind the alveolar ridge
6. Palatal (PL): consonants articulated with the body of the tongue raised against the hard palate
7. Velar (VE): consonants articulated with the back part of the tongue against the soft palate
8. Nasal (NA): consonants produced with a lowered velum, allowing air to escape freely through the nose
9. Stop (SP): consonants in which the vocal tract is blocked, so that all airflow ceases
10. Affricate (AF): consonants that begins as a stop and conclude with a sound of friction

11. Fricative (FR): consonants produced by the friction of breath in a narrow opening
12. Glides (GL): consonants which have is a sound that is phonetically similar to a vowel
13. Liquid (LQ): consonants produced when the tongue approaches a point of articulation within the mouth but does not come close enough to obstruct
14. Vowel (VW): marks that the phoneme is a vowel, without any specifications

Features 2 to 7 specify the place of articulation of the phonemes, while the manner is given by features 8-13. Feature 1 concerns voiced consonants and feature 14 marks the presence of a vowel. Table 7 reports the phonological feature encoding for the phonemes of the word <decadere> (/dekad'ere/) 'to decay'. Although the three phonemes /d/,/k/ and /r/ are different, they share common phonological features as, for example, *Stop* for /d/ and /k/ or *Voicing* and *Dental-alveolar* for /d/ and /r/.

	Phonological features													
Phoneme	VO	BI	LD	DA	PA	PL	VE	NA	SP	AF	FR	GL	LQ	VW
d	1	0	0	1	0	0	0	0	1	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0	0	0	0	0	1
k	0	0	0	0	0	0	1	0	1	0	0	0	0	0
a	0	0	0	0	0	0	0	0	0	0	0	0	0	1
r	1	0	0	1	0	0	0	0	0	0	0	0	1	0
e	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 7. Phonological feature encoding for the phonemes of the word <decadere> (/dekad'ere/) 'to decay', see Table 4.

We designed our test sets by sampling respectively $10,000$ stressed wordforms from GLAFF-IT and absent from PhonItalia, and $10,000$ stressed wordforms from PhonItalia and absent from GLAFF-IT. Our training dataset is composed of all the stressed wordforms reported in Table 6, excluding those used for the test sets. We implemented different architectures for the model by varying the number of neurons of the hidden layer. For a set of architectures, POS information have also been considered as features in the input words. A detailed evaluation of the stress prediction output is reported in the next section.

4 Evaluation and Conclusions

We designed four architectures by varying the size of the hidden layer (5, 10, 20 and 40 neurons). For each architecture, we trained and evaluated two models, including or excluding the POS information of the word. In our approach, a word presents one (and only one) stressed vowel. During the testing phase, we identify the vowel in the word with the highest probability of being stressed: this vowel represents the stressed vowel and thus determines the stress position of the word. Selecting the vowel with the highest probability is computationally convenient because we do not have to identify the probability threshold for separating stressed and unstressed vowels. The Figure 5 displays the

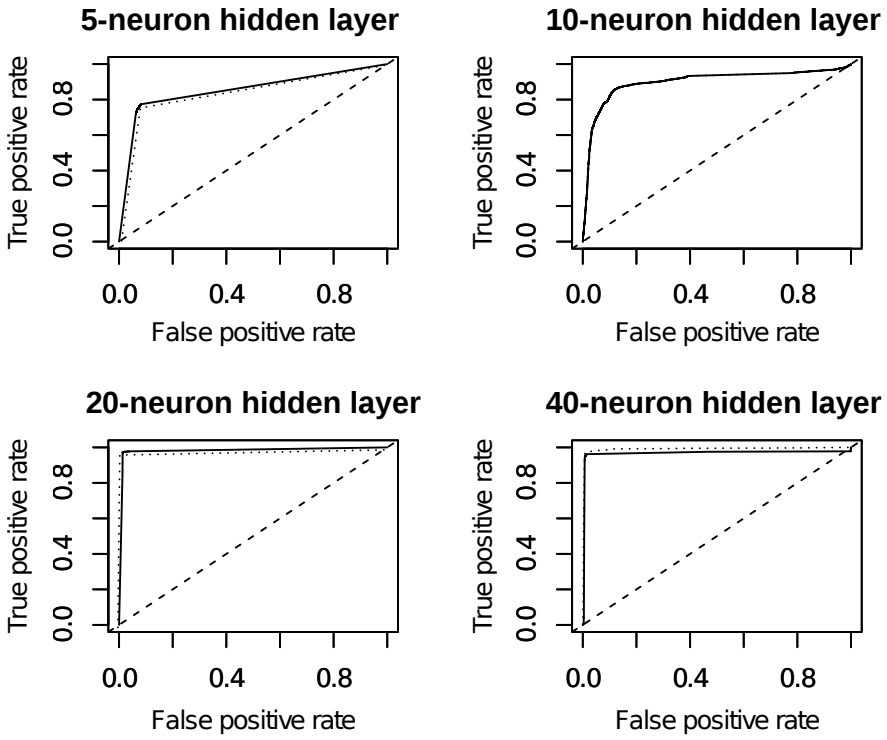

Fig. 5. ROC curves for the four architectures. Testing dataset: $10,000$ stressed wordforms from GLAFF-IT. Solid lines represent models integrating POS information about the word, dotted lines represent models excluding this information.

behaviour of the four architectures by ROC curves with respect to the $10,000$ GLAFF-IT test wordforms. The ROC curves show the trade-off between the true positive rate (sensitivity, *y*-axis) and the false positive rate (1 - specificity, *x*-axis). Sensitivity refers to the proportion of the stressed vowels whereas specificity refers to the proportion of the unstressed vowels. The closer the curve follows the left-hand border, from the origin of axes (0.0 and 0.0) to the top left corner (0.0 and 1.0) and then to the top right corner (1.0 and 1.0), the more accurate the classification. The 5- and 10-neuron models exhibit a substantial false positive rate for the test data, meaning that the model predicts unstressed vowels wrongly categorised as stressed. The 20- and 40-neuron models are very good at classifying with a nearly perfect separation between the stressed and the unstressed vowels. The POS information (dotted lines in Figure 5) does not seem to significantly affect the prediction. The Table 8 reports the percentages of correct stress prediction for the $10,000$ words of the two testing datasets. We observe that the POS information does not improve the prediction (in some cases models without POS information exhibit a better prediction than the models with POS). The grammatical class was a crucial factor for the applicability of the heuristic rules used to predict the stress placement. We can see here that this is not true the stress prediction by the *logit* model.

	Testing - 10,000 words	
Model	GLAFF-IT	PhonItalia
5-POS	73.31%	68.60%
5-NO-POS	71.87%	68.80%
10-POS	84.10%	74.70%
10-NO-POS	83.41%	75.65%
20-POS	**98.12%**	**89.10%**
20-NO-POS	97.89%	88.31%
40-POS	94.75%	82.10%
40-NO-POS	97.18%	87.48%

Table 8. Correct stress prediction for the words in GLAFF-IT and PhonItalia testing dataset.

The 20-neuron model with POS information performs the best prediction and reaches more than 98% of correct stress predicted. We also notice a difference in terms of stress prediction between the two testing sets: the evaluation performed against the PhonItalia test reaches 89.1% of correct predictions by the best model (20-POS). This difference can be explained by the nature of the two test sets. For instance, the PhonItalia test lists a great number of loanwords such as <fairplay>, <academy> or <mission> which have a phonotactic structure totally different from the words of the training dataset. Moreover, the PhonItalia test contains verbs with clitic pronouns as <mangiarlo> 'to eat it' or locative clitic as <andateci> 'you (2nd person plural) go there' which are totally absent from the training data.

In this article, we have described the design of GLAFF-IT, a large-scale morphological and phonological lexicon for Italian language. In order to build this lexicon, we have implemented a set of methods to automatically i) complete inflectional paradigms by generating the missing forms ii) generate missing phonological transcriptions from the orthographic forms and finally iii) predict the stress placement.

The hybrid method we designed for automatic stress prediction, based on a set of heuristic rules and the responses of a *logit* model, reliably predicts stressed and unstressed vowels. Applied to GLAFF-IT, it reaches an accuracy of 98.12%. To our knowledge, GLAFF-IT is the only free Italian lexicon featuring a large coverage ($457,702$ entries) and reporting phonological transcription with stress markings. Moreover, this model will be useful when updating the resource with the new entries regularly added to Wikizionario. Indeed, if contributors are prone to add new entries, they often neglect to provide inflectional and phonological information.

In the near future, we plan to add syllable boundaries to the phonological transcriptions of GLAFF-IT. Regarding the stress prediction, we intend to evaluate the adaptability of our model to other languages presenting variable stress placement. In a psycholinguistic perspective, the model's responses could be compared to the responses provided by speakers with respect to the same set of stimuli, in order to assess the possible correlation between speakers and automatic predictions.

Acknowledgement

Computations performed to produce GLAFF-IT have been carried out using the OSIRIM platform, that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government and ERDF.

References

1. Bafile, L.: Antepenultimate stress in italian and some related dialects: Metrical and prosodic aspects. Rivista di linguistica 11(2), 201–229 (1999)
2. Behravan, H., Hautamäki, V., Siniscalchi, S.M., Kinnunen, T., Lee, C.H.: i-Vector modeling of speech attributes for automatic foreign accent recognition. IEEE/ACM Trans. Audio, Speech & Language Processing 24(1), 29–41 (2016)
3. Bel, N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A.: SIMPLE: A General Framework for the Development of Multilingual Lexicons. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece (2000)
4. Bertinetto, P.M., Burani, C., Laudanna, A., Lucia Marconi, D.R., Rolando, C.: Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS) (2005), `http://linguistica.sns.it/CoLFIS/Home.htm`
5. Calderone, B., Bertinetto, P.M.: From phonotactics to syllables. a psycho - computational approach. In: 46th Annual Meeting of the Societas Linguistica Europaea. Split, Croatia (2013)
6. Calderone, B., Sajous, F., Hathout, N.: GLAW-IT: A free large Italian dictionary encoded in a fine-grained XML format. In: Proceedings of the 49th Annual Meeting of the Societas Linguistica Europaea (SLE 2016). pp. 43–45. Naples, Italy (2016)
7. Dou, Q., Bergsma, S., Jiampojamarn, S., Kondrak, G.: A Ranking Approach to Stress Prediction for Letter-to-phoneme Conversion. In: Proceedings of the 47th Annual Meeting of the ACL. pp. 118–126. Association for Computational Linguistics, Suntec, Singapore (2009)
8. Goslin, J., Galluzzi, C., Romani, C.: PhonItalia: a phonological lexicon for Italian. Behavior Research Methods 46(3), 872–886 (2014)
9. Guion, S.G.: Knowledge of english word stress patterns in early and late korean-english bilinguals. Studies in Second Language Acquisiti 27, 503–533 (2005)
10. Hathout, N., Sajous, F.: Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia (2016)
11. Krämer, M.: The Phonology of Italian. Oxford University Press, Oxford (2009)
12. Peperkamp, S.: Lexical exceptions in stress systems: Arguments from early language acquisition and adult speech perception. Language 80(1), 98–126 (2004)
13. Rajman, M., Lecomte, J., Paroubek, P.: Grace gtr-3-2.1. Tech. rep., EPFL & INaLF (1997)
14. Roventini, A., Alonge, A., Calzolari, N., Magnini, B., Bertagna, F.: ItalWordNet: a Large Semantic Database for Italian. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2000). pp. 783–790. Athens, Greece (2000)
15. Sajous, F., Hathout, N.: GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In: Proceedings of the of the eLex 2015 conference. pp. 405–426. Herstmonceux, England (2015)
16. Slowiaczek, L.M.: Stress and context in auditory word recognition. Journal of Psycholinguistic Research 20(6), 465–481 (1991)
17. Talamo, L., Celata, C., Bertinetto, P.M.: derIvaTario: An Annotated Lexicon of Italian Derivatives. Word Structure 9(1) (2016)
18. Zweigenbaum, P., Grabar, N.: Accenting unknown words in a specialized language. In: Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain. pp. 21–28. Philadelphia, PA (2002)