Verb Morphomes ITA

Morphomic patterns in Italian verbal inflection

https://github.com/franfranz/Verb_morphomes_ITA

1 What's inside Verb Morphomes ITA

This repository contains two lists of inflected verb forms of Italian, Table of Types - `Verb Morphomes ITA Typefreq` and Table of Token Frequency - `Verb Morphomes ITA Tokenfreq`.
The declensional tables contain orthographic and phonetic information on inflected verb forms (*Present Infinitive; Present Idicative 1ps, 3ps, 1 pp; Present Subjunctive 1ps*), orthographic and phonetic information of stem/roots and annotation of subregularities as the presence of morphomic L-patterns; (see section 4 and section 3 for the annotation details).
Token frequency measures for the word forms have been collected from the `Noun` frequency list available from `this repository`. The lists are based on data from ItWaC, a freely available two-billion token corpus obtained from websites in Italian (Baroni et al., 2009).
The morphological tagging for lemma, tense, mood, person, number were obtained from Morph-it!, a morphologically annotated list comprising approximately 500000 Italian word types (Zanchetta and Baroni, 2005). The tag for conjugation class was obtained from the infinitive forms of lemmas.

2 Phonetic transcription

Phonetic forms of consonant graphemes that have more than one pronunciation are transcribed using an adapted version of X-SAMPA:

- S = voiceless postalveolar fricative [ʃ]
- tS / ttS = voiceless alveloar affricate (with postalveolar solution), short/geminate [tʃ] / [tːʃ]
- dG / ddG = voiced alveloar affricate (with postalveolar solution), short/geminate [dʒ]
- N / NN = nasal palatal short/geminate [ɲ]
- L / LL = lateral palatal short/geminate [ʎ]

3 Table of Types - Verb Morphomes ITA Typefreq

In the `Verb Morphomes ITA typefreq` list, 6117 verb lemmas and their inflected forms are ordered in a .csv table, whose columns contain the following information:

- ”lemma_morphit”, infinitive form of the verbal lemma, as reported in the morph-it! list
- ”form_1ps_ind”, present indicative first person singular form, orthographic representation
- ”fonform_1ps_ind”, present indicative first person singular form, phonetic representation
- ”fonroot_1ps_ind”, present indicative first person singular form, phonetic representation of the root/stem
- ”pal_end_1ps_ind”, last phoneme of 1ps_ind is ”tS”,”dG”,”NN”,”LL”,”g”,”k”, or group ”lg”, ”ng”. Other phonemes are not listed.
- ”conj”, conjugation
- ”inf_root”, root of the infinitive form
- ”pal_end_inf”, last phoneme of infinitivr is ”tS”,”dG”,”NN”,”LL”,”g”,”k”, or group ”lg”, ”ng”. Other phonemes are not listed.
- ”form_3ps_ind”, present indicative third person singular form, orthographic representation
- ”fonform_3ps_ind”, present indicative third person singular form, phonetic representation
- ”fonroot_3ps_ind”, present indicative third person singular form, phonetic representation of the root/stem
- ”pal_end_3ps_ind”, last phoneme of 3ps_ind is ”tS”,”dG”,”NN”,”LL”,”g”,”k”, or group ”lg”, ”ng”. Other phonemes are not listed.
- ”form_1pp_ind”, present indicative first person singular form, orthographic representation
- fonform_1pp_ind”, present indicative first person plural form, phonetic representation
- ”fonroot_1pp_ind”, resent indicative first person plural form, phonetic representation of the root/stem
- ”pal_end_1pp_ind”, last phoneme of 1pp_ind is ”tS”,”dG”,”NN”,”LL”,”g”,”k”, or group ”lg”, ”ng”. Other phonemes are not listed.
- ”form_1ps_sub”, present subjunctive first person singular form, orthographic representation
- ”fonform_1ps_sub”, present subjunctive first person singular form, phonetic representation

- ”fonroot_1ps_sub”, present subjunctive first person singular form, phonetic representation of the root/stem
- ”pal_end_1ps_sub”, last phoneme of 1ps_sub is ”tS”,”dG”,”NN”,”LL”,”g”,”k”, or group ”lg”, ”ng”. Other phonemes are not listed.
- ”L_morph”, presence of a L-morphome pattern
- ”indpres_1s_eq_inf”, presence of the same root in the present 1ps_ind and in the present infinitive
- ”fonroot_noinc_1ps_ind” present indicative first person singular form stripped of ”incoative” morpheme (for 3rd conjugation), phonetic representation

4 Table of Token Frequency - Verb Morphomes ITA Tokenfreq

In the `Verb Morphomes ITA typefreq` list, 2933 lemmas and their inflected forms are ordered in a .csv table. Please note that in this table only the types present both in morph-it and in ItWac were listed. The number of verb lemmas is downsized if compared to the Types table. The columns of this table are the same as reported in § 3. Four added columns report the raw count of token frequency as occurring in the ItWac corpus:

- ”formfreq_1ps_ind” Token frequency of the form occurring in the present indicative first person singular and its homographs[*]
- ”formfreq_3ps_ind” Token frequency of the form occurring in the present indicative third person singular and its homographs[*]
- ”formfreq_1pp_ind” Token frequency of the form occurring in the present indicative first person plural and its homographs[*]
- ”formfreq_1ps_sub” Token frequency of the form occurring in the present subjunctive first person singular and its homographs[*]

[*] Some word forms can be homographs or invariant and surface in more than one inflectional feature. Note that the token frequency refers to the total occurrences of the word form, and not to the number of occurrences of the word form when used in a particular inflection.
Some nouns are homograph to other parts of speech, e.g. *canto*, noun-masc.sg. ”song”, or verb-ind-pres-1.sg ”I sing”. The frequency reported here refers only to their occurrence as verbs in the corpus.

Credits

The Verb Morphomes ITA resources are freely available under a GNU General Public license v3.0.
Please acknowledge their use by referring to: Franzon, F. (2021). Morphomic patterns in Italian verbal inflection (Version 4.0.1). https://github.com/franfranz/Verb_morphomes_ITA. Please see repo for citation metadata.
This version is a pre-release. Check the repo for further developments. For any comments or questions, please contact: ffranzon@sissa.it

References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Zanchetta, E. and Baroni, M. (2005). Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics*, 1(1):2005.