Flex It
A large database of Italian inflectional classes.

v1.0.0-b

`https://github.com/franfranz/Flex_it`

1 What's inside Flex It

Flex It is a frequency list containing information on the morphological inflection for 71762 Italian word forms, comprising 33445 noun types and 38317 adjective types.
Each word form is annotated for inflectional class/paradigm, inflectional ending, gender, number, lemma and standardized measures of token frequency, as well as for morphological class-specific information (see section 2 and section 3 for details).
Token frequency measures for the word forms have been collected from the `Noun` and the `Adjective` frequency lists 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, gender, and number were obtained from Morph-it!, a morphologically annotated list comprising approximately 500000 Italian word types (Zanchetta and Baroni, 2005).
The tag for inflectional class was obtained from the inflected forms of lemmas: the methods used to compile the database, and a detailed description of its features are reported HERE() (see sec 3).

2 Flex It — Nouns

In the `Flex It - Nouns` list, the word forms are ordered in a .csv table, whose columns contain the following information:

- "Form", the word form.
- "Lemma", indication for lemma, as occurring in Morph-it!
- "Freq", raw token frequency: the number of times the word form occurs in the ItWaC corpus.
- "Fpmw", token frequency, normalized per million tokens.

- ”Zipf”, token frequency, normalized on a Zipf scale (Van Heuven et al., 2014)
- ”POS”, part of speech (NOUN), tagged for both gender and number (*fp, fs, mp, ms*).
- ”Baseform”, lemma stripped from inflectional suffix; it contains derivational affixes, where present.
- ”Gender”, gender tag: feminine, masculine (*fem. - masc.*).
- ”Number”, number tag: plural, singular (*plur. - sing.*).
- ”Form_amb”, gender and number features in which homographs or invariant forms surface. For instance, the word form *cameriere* ”waiter/waitresses” can be a fem.plur or a masc.sing., and its ”Form_amb” tag reports *fp_ms*. In case the word form is unambiguously related to a unique inflectional feature, *no_amb* is reported.
- ”Ending”, last grapheme (letter) of the word, corresponding to the inflectional suffix in non-invariant words.
- ”Inf_class”, inflectional class, indicated by *ending of singular _ ending of plural*.

Some word forms can be homographs or invariant and surface in more than one inflectional feature. For instance, the word form *cameriere* can be both fem.plur or masc.sing. For these types, all of the forms are listed as unique entries; 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 one of the inflectional features. The type of ambiguity is tagged in the ”Form_amb” column.
Some nouns are homograph to other parts of speech, e.g. *apparecchio*, noun: ”device”, and verb: ”I prepare”. The frequency reported here refers only to their occurrence as nouns in the corpus. In compound nouns such as *capobanda* ”band leader”-sing. *capibanda* ”band leader”-plur., the inflectional morpheme may be transparently marked in the middle of the word. The ”ending” column is not informative on the differences between singular and plurals in these cases.

3 Flex It — Adjectives

In the `Flex It - Adjectives` list, the word forms are ordered in a .csv table, whose columns contain the following information:

- ”Form”, the word form.
- ”Lemma”, indication for lemma, as occurring in Morph-it!
- ”Freq”, raw token frequency: the number of times the word form occurs in the ItWaC corpus.
- ”Fpmw”, token frequency, normalized per million tokens.

- ”Zipf”, token frequency, normalized on a Zipf scale (Van Heuven et al., 2014)
- ”POS”, tag for both gender and number (*fp, fs, mp, ms*).
- ”Baseform”, lemma stripped from inflectional suffix; it contains derivational affixes, where present.
- ”Grade”, grade tag: positive, superlative, comparative (*pos - sup - com*)
- ”Gender”, gender tag: feminine, masculine (*fem. - masc.*).
- ”Number”, number tag: plural, singular (*plur. - sing.*).
- ”Form_amb”, gender and number features in which homographs or invariant forms surface. For instance, the word form *entusiasta* ”enthusiastic” can be a fem.sing or a masc.sing., and its ”Form_amb” tag reports *fs_ms*. In case the word form is unambiguously related to a unique inflectional feature, *no_amb* is reported.
- ”Ending”, last grapheme (letter) of the word, corresponding to the inflectional suffix in non-invariant words.
- ”Inf_class”, inflectional class/paradigm. The inflection of adjectives comprises both gender and number endings. The endings occur in this order: ”ending of singular feminine”_”ending of plural feminine” / ”ending of singular masculine”_”ending of plural masculine”. In case a form is absent in the paradigm or unattested in the corpus, a ”NA” tag is present in its place. In case an adjective displays the same form when inflected in all of the features, a ”Inv” (Invariant) tag is reported.

Some word forms can be homographs or invariant and surface in more than one inflectional feature. For instance, the word form *entusiasta* ”enthusiastic” can be both fem. sing. or masc.sing. For these types, all of the forms are listed as unique entries; 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. The type of ambiguity is tagged in the ”Form_amb” column.
Some nouns are homograph to other parts of speech, e.g. *corte*, adj-fem. plur. ”short”, or noun-fem.sing. ”court”. The frequency reported here refers only to their occurrence as adjectives in the corpus.

Credits

The Flex It resources are freely available. This version of Flex It is a pre-release: please acknowledge its use by citing this repo - a reference paper will be available soon. 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.

Van Heuven, W. J., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

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