10.12871/CLICIT2014123

The New Basic Vocabulary of Italian as a Linguistic Resource

Isabella Chiari
Dip. di Scienze documentarie, linguistico-filologiche e geografiche dell'Università di Roma "La Sapienza"
pl.le Aldo Moro, 5, 00185 Roma
isabella.chiari@uniroma1.it

Tullio De Mauro
Dip. di Scienze documentarie, linguistico-filologiche e geografiche dell'Università di Roma "La Sapienza"
pl.le Aldo Moro, 5, 00185 Roma
tullio.demauro@uniroma1.it

Abstract

English. The New Basic Vocabulary of Italian (NVdB) is a reference linguistic resource for contemporary Italian describing most used and understood words of the language. The paper offers an overview of the objectives of the work, its main features and most relevant linguistic and computational applications.

Italiano. Il Nuovo Vocabolario di Base della lingua italiana (NVdB) costituisce una risorsa linguistica di riferimento dell'italiano contemporaneo che descrive le parole più usate e conosciute dalla maggioranza della popolazione italiana con una istruzione media inferiore. Il contributo descrive le ragioni dell'impianto del NVdB, le caratteristiche della risorsa e le principali applicazioni linguistiche e computazionali.

1 Introduction

Core dictionaries are precious resources that represent the most widely known (in production and reception) lexemes of a language. Among the most significant features characterizing basic vocabulary of a language is the high textual coverage of a small number of lexemes (ranging from 2,000 to 5,000 top ranking words in frequency lists), their large polysemy, their relationship to the oldest lexical heritage of a language, their relevance in fist and second language learning and teaching and as reference tools for lexical analysis.

Many recent corpus based works have been produced to provide up-to-date core dictionaries to many European languages (e.g. the Routledge frequency dictionary series). Italian language has a number reference frequency lists all of which are related to corpora and collections of texts dating 1994 or earlier (among the most relevant Bortolini et al., 1971; Juilland and Traversa, 1973; De Mauro et al., 1993; Bertinetto et al. 2005).

The Basic Vocabulary of Italian (VdB, De Mauro, 1980) first appeared as an annex to *Guida all'uso delle parole* and has been subsequently included in all lexicographic works directed by Tullio De Mauro, with some minor changes.

VdB has benefited from a combination of statistical criteria for the selection of lemmas (both grammatical and content words) mainly based on a frequency list of written Italian, LIF (Bortolini et al., 1972) and later on a frequency list of spoken Italian, LIP (De Mauro et al., 1993) – and independent evaluations further submitted to experimentation on primary school pupils.

The last version of VdB was published in 2007 in an additional tome of GRADIT (De Mauro, 1999) and counts about 6,700 lemmas, organised in three vocabulary ranges.

Fundamental vocabulary (FO) includes the highest frequency words that cover about 90% of all written and spoken text occurrences [*appartamento* 'apartment', *commercio* 'commerce', *cosa* 'thing', *fiore* 'flower', *improvviso* 'sudden', *incontro* 'meeting', *malato* 'ill', *odiare* 'to hate'], while high usage vocabulary (AU) covers about 6% of the subsequent high frequency words [*acciaio* 'steel', *concerto* 'concert', *fase* 'phase', *formica* 'ant', *inaugurazione* 'inauguration', *indovinare* 'to guess', *parroco* 'parish priest', *pettinare* 'to comb']. On the contrary high availability (AD) vocabulary is not based on textual statistical resources but is derived from a psycholinguistic insight experimentally verified, and is to be intended in the tradition of the *vocabulaire de haute disponibilité*, first introduced in the *Français fondamentale* project (Michéa, 1953; Gougenheim, 1964). VdB thus integrates high frequency vocabulary ranges with the so-called high availability vocabulary (*haute disponibilité*) and thus provides a full picture of not only written and spoken usages, but also purely mental usages of word (commonly regarding words having a specific relationship with the concreteness of ordinary life) [*abbaiare* to 'bark', *ago* 'needle', *forchetta 'fork', mancino* 'left-handed', *pala* 'shovel', *pescatore* 'fisherman'].

From the first edition of VdB many things have changed in Italian society and language: Italian language was then used only by 50% of the population. Today Italian is used by 95% of the population. Many things have changed in the conditions of use of the language for the speakers and the relationship between Italian language and dialects have been deeply transformed.

The renovated version of VdB, NVdB (Chiari and De Mauro, in press), will be presented and previewed in this paper. NVdB is a linguistic resource designed to meet three different purposes: a linguistic one, to be intended in both a theoretical and a descriptive sense, an educational-linguistic one and a regulative one, for the development of guidelines in public communication.

The educational objective is focused on providing a resource to develop tools for language teaching and learning, both for first and second language learners. The descriptive lexicological objective is providing a lexical resource that can be used as a reference in evaluating behaviour of lexemes belonging to different text typologies, taking into account the behaviour of different lexemes both from an empirical-corpus based approach and an experimental (intuition based) approach and enable the description of linguistic changes that affected most commonly known words in Italian from the Fifties up to today. The descriptive objective is tightly connected to the possible computational applications of the resource in tools able to process general language and take into account its peculiar behaviour. The regulative objective regards the use of VdB as a reference for the editing of administrative texts, and in general, for easy reading texts.

2 Overview of the resource

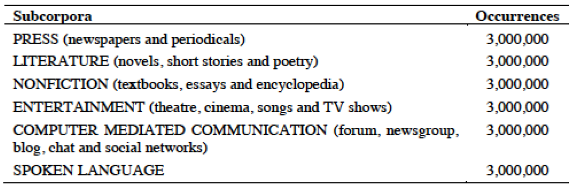
NVdB is characterised by a number of methodological choices that make it a unique tool both for educational, descriptive and computational linguistics. A major feature of NVdB is its stratification in vocabulary ranges. While other lexicographic works contain only a plain list of frequent words, NVdB is organised internally and reveals different statistical and non statistical properties of the elements of the lexicon. The stratification of NVdB, though complex methodologically, allows isolating the different textual behaviour of lexemes in context, their coverage power and dispersion, and also taking into account separately known words that rarely appear in text corpora but that are generally available to native speakers and that necessitate experimental methods to be acquired.

A new experimentation of high availability words completes and redefines the role of frequency and usage introducing a receptive and psycholinguistic perspective in the third layer of the core dictionary.

In order to facilitate applicative uses of NVDB all data will be distributed both in paper and in an open source electronic versions in multiple formats.

2.1 The corpus and linguistic processing

The first two layers of NVdB (FO, AU) are derived by the analysis of a specifically built corpus of contemporary Italian (written and spoken), of 18,000,000 words. The corpus is organized in 6 subcorpora of similar size, further normalized: press (newspapers and periodicals), literature (novels, short stories, poetry), nonfiction (textbooks, essays, encyclopaedia), entertainment (theatre, cinema, songs, and TV shows), computer mediated communication (forum, newsgroup, blog, chat and social networks), spoken language.



Subcorpora	Occurrences
PRESS (newspapers and periodicals)	3,000,000
LITERATURE (novels, short stories and poetry)	3,000,000
NONFICTION (textbooks, essays and encyclopedia)	3,000,000
ENTERTAINMENT (theatre, cinema, songs and TV shows)	3,000,000
COMPUTER MEDIATED COMMUNICATION (forum, newsgroup, blog, chat and social networks)	3,000,000
SPOKEN LANGUAGE	3,000,000

Figure 1: NVdB corpus

The chronological span of the texts included in the corpus range from 2000 to 2012, not diachronically balanced, with a polarization on the last two years. The general criteria for the selection of texts were maximum variability in authors' and speakers' characteristics. Texts produced during the last years were preferred to older ones. For printed materials we have chosen texts from widely known sources (for example using book charts and prize-winners, most read periodicals and TV shows, statistics of blogs and forum access, etc.). As for length, to have to maximize variability of text features we have preferred shorter works over longer ones, always trying to include texts in their integrity.

The corpus has been POS tagged and extensively corrected manually for all entries belonging to the NVdB (Chiari and De Mauro, 2012). POS tagging has been performed in different sessions. The TreeTagger (Schmid, 1994) has been used with the Italian parameter file provided by Marco Baroni as a first step. Errors were corrected and a new reference dictionary has been built in order to perform further correction sessions. Lemmatization procedures and principles were conducted using GRADIT as the main reference tool and thus follows the guidelines of traditional lexicography (De Mauro, 1999). The main consequence of this choice is that VdB does not appear as a flat frequency list in which each line is a couple lemma-POS, but is a hierarchical list o lemmas (as will be discussed in the next paragraph).

Extensive manual correction has involved correction of proper names tagging, of unknown forms, of incorrect lemma-POS attributions, especially regarding high frequency errors. Manual correction has been performed by checking and disambiguating cases in concordances produced for each item in the list (lemma or word form). A special session of lexical homograph disambiguation has been performed fully manually in order to assure complete alignment of the VdB resource results to GRADIT dictionary.

An evaluation of the amount of manual correction of data is fully provided in the documentation.

2.2 Organization of linguistic data in the resource

One of the most significant improvements in the linguistic resource relies on the fact that all data (relative frequency, usage, dispersion) is given in detail for all subcorpora in order to evaluate different behaviour of lexical units in different subcorpora.

The criteria for the organization of the entries in the lexicon follow lexicographic principles and are perfectly aligned to the entries of GRADIT (De Mauro, 1999). Thus while an ordinary frequency list is a flat list of couples represented by the citation form of a lemma and its grammatical qualification (e.g. *cattivo* noun, *fare* verb appear as different entries – and ranks – from *cattivo* adjective and *fare* noun), the internal organization of NVdB is hierarchical: each entry is conceived as a full lexical entry (presumably as saved in the mental lexicon) where each lemma/entry can be associated to more than one grammatical qualification. In NVdB the entry *cattivo* has a general rank, frequency, usage deriving from the sum of its different grammatical realizations, and will also provide detailed information on the frequency and usage of each of the grammatical qualification for the overall corpus and all subcorpora.

Furthermore for the first time in a frequency list and in a core dictionary extensive account of lexical/absolute homographs has been provided (by disambiguating concordance lines manually for all top ranked lemmas). While textual/relative homography is generally addressed in POS tagging, absolute homography is still a significant challenge and cannot be performed adequately by automatic tools. Thus entries in NVdB and their quantitative data make distinction between *riso* noun ('risata', 'alimento'); *calcio* noun ('gioco/pedata', 'elemento chimico'); *asse* noun ('tavola', 'linea…'); *avanzare* verb ('andare avanti', 'essere in sovrabbondanza'); *buono* noun ('il bene o persona buona', 'documento che dà diritto a ricevere un servizio'). Manual disambiguation of lexical homographs touched about 8,3% of all occurrences in the corpus.

Full processing (cumulative and relative) of formal orthographic variants especially needed in case of loanwords (e.g. *goal*, *gol*; *email*, *e-mail*) is provided.

Moreover one of the major novelties in NVdB is the processing and inclusion of multiword expressions (idioms, fixed expressions, named entities) in the lemma list, both marked independently (lemmatised) and cross referenced under main lemma entries (e.g. *al fine di* is a conjunctional idiom lemmatised autonomously and cross-referenced under the headword *fine*). Multiword expressions included in the NVdB follow the main threshold of the AU layer of the general vocabulary list. Data on multiwords belonging to all grammatical categories have been provided by projecting lemmatized version of the reference list of multiwords included in the largest lexicographic work available for Italian

(GRADIT, De Mauro 1999), 67,678 multiword lemmas, also taking into account possible modifiers occurring between multiwords. Data on multiwords has been fully manually checked in order to exclude multiword sequences that are not used idiomatically in the form of fixed expression. Multiwords belonging to the basic vocabulary are provided in a separate lemmatized list.

The final layers of NVdB describe about 7,400 lexemes: about 2,000 fundamental lexemes, about 3,000 high usage lexemes and about 2,400 high availability lexemes.

3 A short overview of data

Interpreting the comparison between VdB and NVdB can be a very difficult task since there are methodological and linguistic (internal) factors that interact inextricably in results. The main problems derive from the different size and design of the two corpora used and internally comparison of lexical differences in usage is insufficient to provide a full interpretation of the new data presented in the NVdB. It is thus capital to merge quantitative and qualitative analysis and to interconnect lexical stableness, shifts and changes to cultural and social changes that occurred in Italy in the past fifty years.

A rough snapshot of the stableness of Vdb (1980) can be seen by observing how much of the old layers still belongs to the same layer in NVdB. 73.3% of the old FO is stable, 47% of AU is preserved. Most new entries in the new FO layer previously belonged to AU (15% of the overall new FO). Examples of AU lexemes that migrated to FO layer are: *adulto* 'adult', *anziano* 'old' ', *assenza* 'absence', *camion* 'truck', *buco* 'hole', *cassa* 'box', *codice* 'code', *concerto* 'concert', *individuo* 'individual', *insegnante* 'teacher', *lavoratore* 'worker', *letteratura* 'literature', *maggioranza* 'majority', *paziente* 'patient', *procedura* 'procedure', *reagire* 'to react', *ruolo* 'role', *ritmo* 'rhythm', *strumento* 'instrument', *telefonata* 'phone call', *turno* 'turn'.

Other words dropped from FO: *aggiustare* 'to fix', *agricoltura* 'agriculture', *animo* 'soul', *calma* 'calmness', *carità* 'charity', *collina* 'hill', *cretino* 'idiot', *ebbene*, *educare* 'to educate', *fidanzato* 'fiancée', *guaio* 'trouble', *illuminare* 'to illuminate', *ladro* 'thief', *mela* 'apple', *noioso* 'boring', *occupazione* 'occupation', *patria* 'homeland', *pietà* 'pity', *provinciale* 'provincial', *valigia* 'suitcase', *vasto* 'wide', ^{1}*volgare* 'vulgar'.

4 Conclusion and future developments

The NVdB of Italian is be distributed as a frequency dictionary of lemmatized lexemes and multiword, with data on coverage, frequency, dispersion, usage labels, grammatical qualifications in all subcorpora. A linguistic analysis and comparison with previous data is also provided with full methodological documentation.

The core dictionary and data are also distributed electronically in various formats in order to be used as a reference tool for different applications.

Future work will be to integrate data from the core dictionary with new lexicographic entries (glosses, examples, collocations) in order to provide a tool useful both for first and second language learners and for further computational applications.

References

Umberta Bortolini, Carlo Tagliavini and Antonio Zampolli, 1971. *Lessico di frequenza della lingua italiana contemporanea*, IBM Italia, Milano.

Isabella Chiari and Tullio De Mauro. 2012. The new basic vocabulary of Italian: problems and methods, in «Statistica applicata», 22 (1), 2012, pp. 21-35.

Isabella Chiari and Tullio De Mauro. In press. *Il Nuovo Vocabolario di Base della lingua italiana*. Casa Editrice Sapienza, Roma.

Tullio De Mauro. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.

Tullio De Mauro. 1999. *Grande Dizionario Italiano dell'uso.* UTET, Torino.

Tullio De Mauro, Federico Mancini, Massimo Vedovelli and Miriam Voghera, 1993. *Lessico di frequenza dell'italiano parlato (LIP)*, Etas libri, Milano.

Georges Gougenheim. 1964. *L'élaboration du francais fondamental (1er degré): Étude sur l'établissement d'un vocabulaire et d'une grammaire du base*. Didier, Paris.

Juilland, Alphonse G. and Vincenzo Traversa. 1973. *Frequency Dictionary of Italian Words*, Mouton, The Hague,.

René Michéa. 1953. Mots fréquents et mots disponibles. Un aspect nouveau de la statistique du langage. *Les langues modernes*. (47): 338-44.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Paper presented to the Proceedings of International Conference on New Methods in Language Processing*: 44-49.