THE NEW BASIC VOCABULARY OF ITALIAN: PROBLEMS AND METHODS

Isabella Chiari, Tullio De Mauro[1]

Department of Document Studies, Linguistics and Geography, University "La Sapienza", Rome, Italy

Abstract*. In this contribution we will introduce some of the methodological and scientific innovations introduced in the project for the new basic vocabulary of Italian. A brief overview of previous versions of the basic vocabulary and of its main characteristics will be specified and a presentation of the key methodological innovations in the new project, developed about thirty years later will follow. We will present some of the central choices, features and problems encountered in corpus building and processing, and present some of the most significant linguistic aspects that the new version of the basic vocabulary will bear.*

Keywords*: Italian, Basic vocabulary, Corpus linguistics, Linguistic resources*

1. INTRODUCTION

The basic vocabulary of Italian (VDB)[2] was designed and released in 1980. We are approaching a renovated version that will be presented and previewed in this paper. VDB is a linguistic resource designed to meet three different purposes: a linguistic one, to be intended in both a theoretical and a descriptive

[1] Isabella Chiari, email: isabella.chiari@uniroma1.it Tullio De Mauro, email: tullio.demauro@uniroma1.it. The present paper is the result of a research project designed and conducted jointly by the two authors. The specific drafting and writing of Section 1 is by Tullio De Mauro, while Sections 1.1,1.2, 2 and 3 have been written by Isabella Chiari.

[2] The acronyms and abbreviations used in this paper follow the common form used for Italian language, in order not to confuse Italian readers familiar with those abbreviated expressions.

sense, an educational-linguistic one and a regulative one, having to do with the development of guidelines in public communication.

Linguistic purpose. Since the 1930s, frequency properties in word occurrence distribution, both in spoken and written texts, have been a focal point in linguistic research. As is now common knowledge, once a frequency word list is established for any given language, only a few words, generally empty grammatical words, cover about half covers the occurrences of any text. About the first 2,000 most frequent words cover an average of 90% of the occurrences. A slightly larger number of words (about 3,000) cover another 5-6%. The rest of the occurrences are covered by the remaining words occurring in texts and spoken conversations. The full acceptance of this framework has led some researchers to pose different questions. The major question, still insufficiently explored, is that of estimating the number of words a specific language holds. The answer is subject to significant differences. Languages mostly used in audio-oral mode in small ethnic and geographical ranges by populations, generally present a lexicon of tens of thousand words. Languages often referred to as culture languages or civilisation languages, namely languages used both in oral and written form in extensive ranges, with a population organised in complex systems and socio-economic, technical, productive and intellectual stratification, present a lexicon bearing an extension incomparably larger.

Up to the 1920s estimates were made referring to the size of the lemma list of large printed dictionaries, such as the Oxford English Dictionary, reaching four or five hundred thousand words. But as observed in the preface to the first edition of *Grande Dizionario Italiano dell'Uso*, GRADIT (De Mauro, 1999) words not occasionally present in texts pertaining to languages used by a population with a complex self articulation can be estimated to millions, and even tens of millions, continuously growing under the pressure of scientific taxonomical lexicon development.

These estimates suggest that a systemic portrait and description of the lexicon of languages is largely inadequate. We must therefore regard the lexicon of languages not as a system that constrains individual usages, *paroles* and speech acts, but as a result of the convergence of different speakers' behaviour. The extent of lexical elements of a language appears massive and intrinsically shifting. The few thousand words of very high frequency represent

a core of the lexicon, which appears relatively stable under the diatopic, diaphasic and also diachronic point of view.

The basic vocabulary, direct heir of frequency word books, characterises the core of the lexical mass of a language. For Italian language the basic vocabulary has shown the diachronic stability of the core of the lexicon and also the role played by the vocabulary of *Divina Commedia* in the construction of this very peculiar layer of Italian lexicon.

VDB 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 words (commonly regarding words having a specific relationship with the concreteness of ordinary life). The basic vocabulary brings us to the heart of culture of a nation in its anthropological sense.

Educational purpose. The oldest frequency dictionaries were driven by the aim of enhancing foreign language teaching by making learners start from the study of high frequency words. Paradoxically, from the same educational perspective, came also the harshest critique: frequency dictionaries omitted words that were perceived as equally important to frequent ones. For example, for Italian they included *coltello* and *cucchiaio*, but not *forchetta*, they documented *scarpa* but not *laccio*. From these consideration 'intuition based' dictionaries for lexical learning were developed.

VDB assumes this critique and overcomes it by integrating data from frequency lists with data from research on high availability words. It thus becomes a comprehensive tool for learning (and teaching) the lexicon of Italian as a second language and for the evaluation of the lexicon acquisition and mastering by pupils of primary school. The full mastering of the portion of lexicon included in VDB is a reasonable educational objective for students after eight years of basic education. A number of surveys show that this portion of vocabulary is generally known to those having at least a junior high school degree (Ferreri 2005). The higher the presence of VDB in texts the higher the possibility that they can be understood by the population having that degree (today about half of the population above fourteen years of age).

Regulative purpose. Following a different path, that of brevity of words used in texts, readability measures determine and evaluate higher or lower degrees of readability depending ultimately on basic vocabulary usage. These were the reasons that led to the use of VDB in the 1980s or the compilation and

editing of the volumes of the series "Libri di base", and further taken as a reference for the editing of administrative texts, and in general, for easy reading texts (Piemontese, 1996; Fioritto, 1997).

Thirty years have passed from the creation of the first VDB. Italian was then used only by half of the population, now is commonly used by 95% of it. Significant changes have occurred in language usage. These reasons already motivated the need of a new revision of VDB. But in the meantime changes have also occurred in methodologies of textual analysis. VDB inherited from frequency dictionaries some rough means of grammatical coding and lexeme disambiguation. The knowledge and practice with *Lessico di Frequenza dell'italiano parlato,* LIP (De Mauro *et al*., 1993) and GRADIT (De Mauro, 1999) have lead to a more detailed and precise coding scheme. Furthermore, since the publication of LIP, the access to spoken language transcription corpora has opened up. Progress in digitalisation and automatic lemmatisation shaped new conditions for the founding of VDB on a far larger textual base than that available in 1970's and 1980's.

This was the background that stimulated the enterprise of a new revised basic vocabulary of Italian to be conceived as a reference tool and a linguistic resource.

1.1 SOME HISTORICAL REMARKS

Previous works with similar aims have been developed mainly as teaching tools and exhibit different approaches and methodologies in the selection of basic lemmas: Thompson (1927), a MA thesis containing a 500 word list; Knease (1933), a selection of about 2,000 words from literary texts; Skinner (1935), a 3,000 word list extracted from textbook of Italian for foreigners; Russo (1947), a combined list of 3,137 words; Migliorini (1943), a collection of 1,500 lemmas selected following the linguist's intuition rather than statistical data. Specifically aimed at selecting fundamental vocabulary are also two works that appeared in the 1970s (Juilland and Traversa, 1973; Sciarone, 1977).

The Basic Vocabulary of Italian (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: DIB (De Mauro *et al*., 1996), DAIC (De Mauro and Cattaneo, 1997), GRADIT (De Mauro, 1999), and De Mauro (De Mauro, 2000).

Following these lexicographic works other dictionaries of Italian have introduced tags indicating fundamental or highly available words ranging from 1,000 to 10,000 entries (e.g. Sabatini-Coletti, Zingarelli, Devoto-Oli). Differences in the number of entries and in the terminology used to describe it is to be assigned to different ideal users and mostly to dissimilarities in selection methods, not always explicitly documented.

1.2 THE BASIC VOCABULARY OF ITALIAN LANGUAGE

VDB, published in 1980, is radically different from similar works having the same main aims, methodology and applications (De Mauro, 1980). The idea of gathering data on usage and availability of words (lemmas) has a long tradition, which was widely developed mainly in French lexicological works that blended statistical methodology and educational and descriptive goals (Michéa, 1949; Michéa, 1953; Gougenheim, 1955; Guiraud, 1960; Gougenheim, 1964; Juilland *et al*., 1970; Juilland and Traversa, 1973; Michéa, 1974).

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 subjected to experimentation on primary school pupils.

The last version of VDB was published in 2007 in an additional tome of GRADIT and counts about 6,700 lemmas, organised in three vocabulary ranges (Table 1).

Table 1. VDB composition (Gradit 1999-2007)

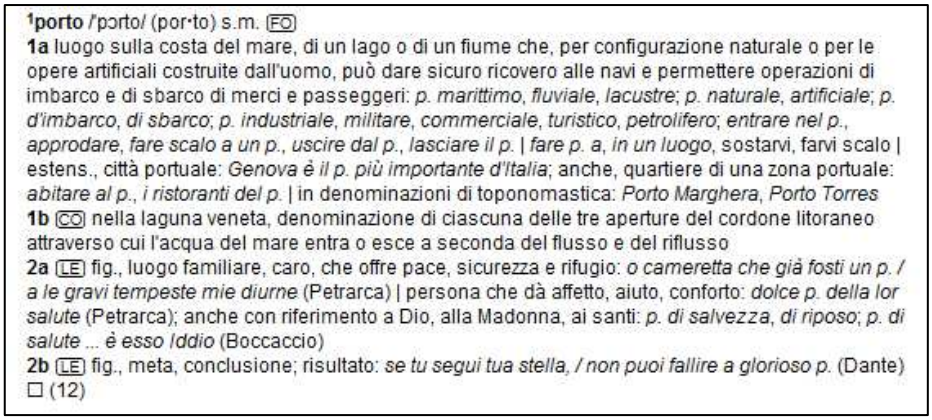
Range	Lexemes
FO – fundamental vocabulary	2,077
AU – high usage	2,663
AD – high availability	1,988

Fundamental vocabulary (FO) includes the highest frequency words that cover about 90% of all written and spoken text occurrences, while high usage vocabulary (AU) covers about 6% of the subsequent high frequency words. 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).

Vocabulary ranges are indicated as *usage marks* or tags that are attached not only to each lemma in the dictionary, but also characterise each word sense belonging to that lemma (see Figure 1).

GRADIT (1999) first introduced a further vocabulary range based on statistical principles, that of *common vocabulary* (CO), comprising about 50,000 words belonging to different disciplines and areas generally known to people having secondary education (De Mauro, 2005: 60).



[1]porto /'pɔrto/ (por·to) s.m. [FO]
1a luogo sulla costa del mare, di un lago o di un fiume che, per configurazione naturale o per le opere artificiali costruite dall'uomo, può dare sicuro ricovero alle navi e permettere operazioni di imbarco e di sbarco di merci e passeggeri: *p. marittimo, fluviale, lacustre*; *p. naturale, artificiale*; *p. d'imbarco, di sbarco*; *p. industriale, militare, commerciale, turistico, petrolifero*; *entrare nel p., approdare, fare scalo a un p., uscire dal p., lasciare il p.* | *fare p. a, in un luogo*, sostarvi, farvi scalo | estens., città portuale: *Genova è il p. più importante d'Italia*; anche, quartiere di una zona portuale: *abitare al p., i ristoranti del p.* | in denominazioni di toponomastica: *Porto Marghera, Porto Torres*
1b [CO] nella laguna veneta, denominazione di ciascuna delle tre aperture del cordone litoraneo attraverso cui l'acqua del mare entra o esce a seconda del flusso e del riflusso
2a [LE] fig., luogo familiare, caro, che offre pace, sicurezza e rifugio: *o cameretta che già fosti un p. / a le gravi tempeste mie diurne* (Petrarca) | persona che dà affetto, aiuto, conforto: *dolce p. della lor salute* (Petrarca); anche con riferimento a Dio, alla Madonna, ai santi: *p. di salvezza, di riposo*; *p. di salute ... è esso Iddio* (Boccaccio)
2b [LE] fig., meta, conclusione; risultato: *se tu segui tua stella, / non puoi fallire a glorioso p.* (Dante)
☐ (12)

Figure 1. Example of usage tags and word senses in GRADIT

VDB from its first publication has been conceived both as a linguistic resource describing the lexicon and as a learning and teaching tool (for first language, L1, and second language, L2). This double objective distinguishes it from similar works, and has generated a large number of applications developed from the 1980s on: the books series of *Libri di Base* (Editori Riuniti); a popular magazine, *Due parole*; measures and indexes of readability for Italian; the integration in a number of lexicographic works conceived for different school levels (De Mauro *et al*., 1996; De Mauro and Cattaneo, 1997) and in the largest dictionary of the Italian language published so far, GRADIT.

VDB is characterised by a number of methodological choices that make it a unique tool both for educational and descriptive linguistics. A major feature of VDB is its stratification in vocabulary ranges. While other lexicographic works contain only a plain list of frequent words, VDB is organised internally and reveals different statistical and non statistical properties of the elements of the lexicon. The stratification of VDB, though complex methodologically, allows isolating the different textual behavior 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 further relevant feature of VDB is its explicitness and transparency in documenting methodological choices and regularity in the application of usage tags and in the collection of textual materials for data extraction (De Mauro, 2005). This requisite is capital for further employment of VDB in the development of other linguistic applications and resources.

2. THE NEW BASIC VOCABULARY OF ITALIAN LANGUAGE

The overall design of VDB has remained basically the same since its first publication in 1980. Some changes were introduced after the arrival of LIP (De Mauro et al. 1993), including data from spoken language corpora, and minor changes have been made in the following editions of GRADIT (from 1999 to 2007).

The need for a revised edition of VDB is based upon three aspects that are relevant from the empirical and methodological point of view. The first aspect is the need for the exploration of important changes that might have occurred and did in fact occur in the previous thirty years: new words have emerged significantly and some have lessened their impact in texts. The aim of the new VDB (NVDB) is the evaluation of the impact of lexical changes in Italian to the core of the vocabulary and the monitoring of shifts from one vocabulary range to another, especially regarding the AD range.

The second relevant aspect is the availability of new computer science tools for (semi-automatic) linguistic data processing and larger data capabilities. The availability of this software makes it possible to perform operations that were previously extremely time-consuming and that enable us to

take into account multiword expressions, homonyms, grammatical properties of lexemes, etc. in a more comprehensive way.

The last aspect regards the applicative power of linguistics resources such as VDB: the need of open source resources that can serve computational linguistics applications and information technology.

2.1 THE NVDB ARCHITECTURE

The NVDB will be characterised by transparency in methodological choices and by open source distribution of all data in multiple formats. From the methodological point of view, the novelties include: a reordering of the main vocabulary ranges (FO, AU, AD) and a reorganisation of their relationships in a bottom-up approach deriving selection and number of items from empirical data; a rethinking of the AD vocabulary taking into account new ways of identifying and experimenting the section of vocabulary that tends not to exhibit high frequency in corpora but results widely known by native speakers; a revision of the lemma list structure; the creation of a corpus of about 18,000,000 running words organised in 6 subcorpora (see Section 2.2:2.2 The NVDB corpus); an explicit marking of quantitative dominant presence in specific subcorpora together with dispersion values; the introduction of a semi-automatic lemmatisation procedure; the marking of specific (multiple) grammatical categories attached to each lemma with relative frequencies (e.g. *cattivo*, with adjective and substantive usages in corpora); the full documentation of all corpus and experimental processing procedures and of criteria for the identification of each vocabulary range; an overall detailed description of the properties of the NVDB and comparative analysis of NVDB and of previous versions of VDB (De Mauro, 1980; De Mauro, 1999).

Relevant novelties in VDB content regard detailed data on raw and relative frequencies of lemmas and forms, dispersion and usage data. Each lemma is accompanied by overall data and frequency on grammatical categories represented in the corpus. Full processing (cumulative and relative) of homographs and formal variants especially needed in case of loanwords (e.g. *goal, gol; email, e-mail*). 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*).

In order to facilitate applicative uses of NVDB all data will be distributed both in paper and electronic versions in multiple formats (txt, xml, xls). Additionally an open source application for the evaluation of basic vocabulary and readability measures in texts provided by the used will be made available online.

2.2 THE NVDB CORPUS

The NVDB corpus is composed of about 18,000,000 word occurrences, organised in 6 subcorpora, covering written (15 millions) and spoken (3 millions) language. The chronological span of the selected texts ranges from 2001 to 2011, even though the corpus has not been diachronically balanced. Since the corpus is not meant to be published as it is, texts covered by copyright issues have nevertheless been included, especially if particularly representative of a specific typology.

Table 2. NVDB corpus composition

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

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.

2.3 TEXT PROCESSING AND DATA EXTRACTION METHODS

Corpus processing was handled in six phases: 1) Corpus pre-processing and text cleaning; 2) Grammatical tagging and lemmatisation; 3) (Manual) error correction and separate processing for lexemes, proper names, numbers and unrecognised forms; 4) Multiword expression extraction and evaluation; 5) Generation of a synoptic table for corpus comparison and dispersion measurement; 6) AD vocabulary evaluation.

Corpus pre-processing and cleaning serves the purpose of making the maximum number of lexemes recognised by POS tagging in the following stages of the procedure. Corpus cleaning is performed with *ad hoc* procedures that depend on special characteristics of the source of textual materials. Specific tools (such as TextSoap, TaLTaC2, and other text processing tools), manually supervised, were employed to overcome encoding problems and irrelevant xml or html tags present in internet captured texts. Some of the most common operations have been: encoding check and conversion; HTML code extraction, URL link eliminations; specific graphic and textual conventions elimination (such as colophon, timing information for subtitles, etc.); accent and apostrophes check and standardisation; lowering of capital letters following strong separators and "all caps" forms; and elimination of non standard characters and of multimedia objects.

Grammatical tagging and lemmatisation has been lead using a probabilistic tagger, TreeTagger (Schmid 1994), with Marco Baroni tagset[3]. The results of lemmatisation has been further imported into the Italian software TaLTaC2 (Bolasco, 2010; www.taltac.it) for text processing and concordance checks.

The following procedure has been adopted in order to provide for each subcorpus the following tables: lemmatised frequency list (where each lemma can occur in multiple grammatical categories associations, e.g. *bello* Noun, *bello* Adjective); word form list (with an associated lemma, a grammatical category and a relative frequency); and a list of all forms that have not been recognised in the lemmatisation procedure and that need manual processing; a list of named entities; a list of all cardinal numbers occurring in digits.

Since no automatic lemmatisation tool is exempt from errors, a specific stage was dedicated to manual check and error correction. Errors belong to the

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/;
ftp://ftp.ims.uni-stuttgart.de/pub/corpora/italian-par-linux-3.2-utf8.bin.gz

following typologies, each having a different correction procedure: a) forms that are categorised as *unknown* by TreeTagger; b) TreeTagger errors in categorisation and/or lemmatisation; c) forms that are not disambiguated by TreeTagger but are tagged with multiple lemmas; and d) absolute homographs disambiguation.

The unknown forms have been extracted from each corpus and manually corrected checking their use in concordances (using TaLTaC2). The amount of forms not recognised by the lemmatiser is generally about 1/6 of all forms (e.g. on a 3,000,000 press subcorpus, about 500,000 occurrences are not recognised, belonging to 96,000 word types – 59,000 of which are hapax). Thus manual correction have been conducted on all unrecognised word types above the threshold of 6 occurrences in each subcorpus. The most common errors in this category are forms belonging to new lemmas, especially acronyms and abbreviations (*www*, *copyright*, *art.*, *HD)*, forms belonging to existing lemmas not recognised for the presence of typos, error in letter case, errors in grammatical attribution, errors in the interpretation of separators' role in context (*l'*, *dell'*), or missing proper nouns categorisation (*Twitter*, *Obama*).

The second error typology is the mismatch of grammatical category and/or lemma attribution. To address the error problem we have chosen to manually check and correct at most the top frequent 10,000 lexical units identified and to further check only items that were above that threshold in at least one subcorpus. The check has been lead on a sampling base (through concordances) and in manually revising the lemmatised list for non existent forms and possible inaccuracies and mistakes (e.g. *orare*, *vociare*, *ombrare*, *idrogenare)*.

Forms that are ambiguously tagged by TreeTagger (e.g. the form *conti* – from *conte* or *conto*, the form *parti* – from *parto* or *parte* are labeled with *conto*|*conte* and *parto*|*parte* tags by TreeTagger) have been fully disambiguated using concordance check.

The absolute homographs issue has been resolved by extracting a list of possible homographs of this kind (belonging to the same grammatical categories and thus not disjointed in the TreeTagger list, such as *riso* as a noun can both refer to 'laughs' and to 'rice') from GRADIT and manually fully disambiguating all forms using concordance check. A further error estimation will be provided using sampling check on existing final data.

Table 3. Example of synoptic lemma table

Lemma	CAT	PRESS	LITER.	NONFIC.	ENTERT.	CMC	SPOKEN	F.TOT	U (F*D)
NON		**27,095.90**	**46,272.24**	**18,529.53**	**71,698.72**	**58,535.58**	**55,045.72**	**277,177.69**	**194,223.07**
	AVV	27,095.9	46,272.24	18,529.53	71,698.72	58,535.58	55,045.72	277,177.69	194,223.07
QUESTO		**105,20.91**	**9,778.91**	**14,135.25**	**20,632.1**	**17,587.42**	**35,279.13**	**107,933.73**	**72,339.36**
	A	7,694.14	6,185.17	11,040.09	11,354.4	11,937.8	22,500.11	70,711.71	48,678.06
	PRON	2,826.77	3,593.74	3,095.17	9,277.7	5,649.62	12,779.02	37,222.02	23,661.29
DIRE		**6,688.76**	**17,253.33**	**4,564.1**	**22,650.6**	**14,509.04**	**26,598.91**	**92,264.74**	**61,288.93**
	V	6,688.76	17,253.33	4,564.1	22,650.6	14,509.04	26,598.91	92,264.74	61,288.93
POTERE[1]		**8,844.13**	**9,660.35**	**9,170.46**	**19,896.76**	**13,629.29**	**16,732.54**	**77,933.52**	**57,232.62**
	V	8,844.13	9,660.35	9,170.46	19,896.76	13,629.29	16,732.54	77,933.52	57,232.62
DOPO		**4,303.85**	**4,356.92**	**2,727.35**	**2,860.08**	**3,525.22**	**4,184.38**	**21,957.79**	**16,539.52**
	AVV	510.72	1,054.11	346.69	864.67	505.59	1,446.87	4,728.66	3,193.62
	CONG	459.08	543.99	302.86	338.11	388.08	641.84	2,673.96	2,029.81
	PREP	3,329.46	2,758.82	2,077.79	1,653.6	2,628.45	2,095.67	14,543.79	11,310.09
PICCOLO		**1,861.56**	**2,565.53**	**1,810.15**	**2,054.53**	**1,603.35**	**1,954.37**	**11,849.49**	**9,623.68**
	A	1,767.45	2,509.74	1,760.84	2,014.8	1,543.05	1,893.67	11,489.55	9,353.08
	N	94.11	55.79	49.3	39.72	60.3	60.7	359.93	270.59
MONDO		**3,277.81**	**1,828.25**	**2,052.75**	**1,700.71**	**2,142.96**	**1,217.00**	**12,219.49**	**9,038.7**
	N	3,277.81	1,828.25	2,052.75	1,700.71	2,142.96	1,217.00	12,219.49	9,038.7

Multiword expression procedures have not yet been performed on existing subcorpora but will include both comparison with existing lists extracted from GRADIT.

Further AD vocabulary experimentation will be performed after the conclusion of NVDB processing of the statistically based vocabulary ranges (FO and AU) and will not be illustrated in the present paper.

After the processing procedure the main resource is represented by a set synoptic tables that give account of the occurrence of each lemma in different grammatical categories and in different subcorpora (with raw and normalised frequency data on all grammatical and content words, see Table 3 for an example).

3. A BRIEF LOOK ON PROVISIONAL DATA

The work on the NVDB is still in progress at the moment of writing, even if procedures for corpus processing have been fully defined and applied already to four out of the six existing subcorpora (press, literature, entertainment and non fiction). We have a perception of some relevant changes in the structure of some of the vocabulary ranges compared to those of the VDB in previous versions.

Some of the most significant changes regard forms that shifted from AU (high usage) vocabulary range and CO (common vocabulary) to the core of VDB, that is the fundamental vocabulary.

For example, lemmas that belonged to AU and are estimated to be shifted to FO are: *controllo, uscita, giudice, messaggio, sesso, arrivo, coppia, nonno, serata, attacco, telefonata, pubblico, divano, televisione, barca, giornalista, turno, disegnare, locale (Adj.), provenire, ignorare, nota, cassa, conversazione, dettaglio, schermo, essere (N.), reagire, pantalone, aereo, sollevare, costume*. While as for shifting from CO to FO we observe lemmas such as: *auto, foto, aperto (Adj.), dollaro, solito (N.), chiuso, scorso, ovunque, interessante, precedente, nascosto, disperato, acceso, armato, evento, villaggio, percorso (N.), pulito, star, complicato, telefonino*. Shifting from common vocabulary to FO are generally many adverbs in *–mente*: *finalmente, probabilmente, veramente, naturalmente, completamente, semplicemente, sicuramente, direttamente, esattamente, ovviamente, certamente, assolutamente.* Also from CO are moving toward the fundamental vocabulary some technical lemmas such as *euro, cellulare, test, manager, video*.

Completely absent or present only in AD or in spoken corpora is a number of words that belong to foul language that now appear well dispersed even in more formal text typologies (e.g. *cazzo, stronzo, merda, coglione, casino, fottere, incazzarsi, cazzata, cesso, pisciare,* and *stronzata*). From the range of high availability (AD) a number of lemmas have showed a tendency in higher occurrence rate and have joined the FO range: *corsa, persino, computer, morto, particolare (N.), stretto, paziente (N.), diamante, chilo, confuso, salone,* and *busta*.

To have a full perspective on changes in vocabulary usage, on the relationship among different vocabulary ranges and on methodological aspects regarding range identification still some research has to be completed. Mainly,

research is needed on the outcomes of the subcorpora processing, multiword expression extraction and AD experimentation.

REFERENCES

Bolasco, S. (2010). *Taltac2.10 Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*, LED, Milano. (http://www.ledonline.it/Taltac/allegati/459-7-TALTAC-COL.pdf)

Bortolini, U., Tagliavini, C. and Zampolli, A. (1972). *Lessico di frequenza della lingua italiana contemporanea*. Garzanti, Milano.

De Mauro, T. (1980). *Guida all'uso delle parole: parlare e scrivere semplice e preciso per capire e farsi capire*. Editori Riuniti, Roma.

De Mauro, T. and Cattaneo, A. (1997). *DAIC: Dizionario avanzato dell'italiano corrente*. Paravia, Torino.

De Mauro, T., Moroni, G.G. and Cattaneo, A. (1996). *DIB: dizionario di base della lingua italiana*. Paravia, Torino.

De Mauro, T. (1999). *Grande Dizionario Italiano dell'uso*. UTET, Torino.

De Mauro, T. (2000). *Il dizionario della lingua italiana*. Paravia, Torino.

De Mauro, T. (2005). *La fabbrica delle parole: il lessico e problemi di lessicologia*. UTET, Torino.

De Mauro, T, Mancini, F., Vedovelli, M. and Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato (LIP)*. Etaslibri, Milano.

Ferreri, S. (2005). *L'alfabetizzazione lessicale: studi di linguistica educativa*. Aracne, Roma.

Fioritto, A. (1997). *Manuale di stile: strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Il Mulino. Presidenza del Consiglio dei Ministri, Milano.

Gougenheim, G. (1955). Le français élémentaire. *International Review of Education/Internationale Zeitschrift für Erziehungswissenschaft/Revue internationale de l'éducation.* (1): 401-412.

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

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Presses Universitaires de France, Paris.

Juilland, A.G., Brodin, D.R. and Davidovitch, C. (1970). *Frequency Dictionary of French Words*. Mouton, The Hague.

Juilland, A.G. and Traversa, V. (1973). *Frequency Dictionary of Italian Words.* Mouton, The Hague.

Knease, T.M. (1933). *An Italian Word List from Literary Sources.* The University of Toronto Press, Toronto.

Michéa, R. (1949). Introduction practique a une statistique du langage. *Les Langues Modernes.* (43): 173-86.

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

Michéa, R. (1974). Les vocabulaires fondamentaux dans l'enseignement des langues vivantes. *Francais dans le Monde.* (13): 11-13.

Migliorini, B. (1943). *Der grundlegende Wortschatz des Italienischen*. Elwert, Marburg.

Piemontese, M.E. (1996). *Capire e farsi capire: teorie e tecniche della scrittura controllata*. Tecnodid, Bari.

Russo, G.A. (1947). A combined Italian word list. *The Modern Language Journal.* (31): 218-40.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*: 44-49. Manchester, UK

Sciarone, A.G. (1977). *Vocabolario fondamentale della lingua italiana*. Guerra, Perugia.

Skinner, L.H. (1935). A comparative study of the vocabularies of forty-five Italian textbooks. *The Modern Language Journal,* (20): 67-84.

Thompson, M.E. (1927). *A Study in Italian Vocabulary Frequency*. Masters' Thesis. University of Iowa, Iowa City.