Multilexical units and headword status. A problematic issue in recent Italian lexicography

Carla Marello
Universita di Torino Italia

The paper will discuss the headword status of multilexical units in Italian monolingual dictionaries and will include a comparison of Italian and Spanish dictionaries. Twentieth century monolingual lexicographies of Romance languages recognized and registered multiword units, but did not promote them easily to headword status. Italian and Spanish monolingual lexicography in particular have very few multilexical units whereas French has a few more. The initial infiltrations through the 'one-word headword' wall came through Latin borrowings (alter ego 'second self', aut aut, 'forced choice', tabula rasa 'blank sheet'), through two (or more for French) centuries of French and Anglo-American multiword borrowings entering gradually into the Italian language and then into monolingual dictionaries macrostructures (for instance ballon d'essai 'trial balloon', malgré lui, 'despite him', fair play, self-made man are XIX century borrowings; j'accuse, 'denunciation', au pair, best seller, on the road are XX century borrowings), and in recent decades through the macrostructures of bilingual English-Italian dictionaries where English multilexical headwords are registered and brought to the attention of Italian monolingual lexicographers as multiword units with headword status in English monolingual dictionaries. A status which might determine them becoming multilexical headwords also in Italian monolingual dictionaries. Nowadays most Italian multiwords still remain registered under one-word headwords, even adjectival or adverbial phrases which cannot occur as single words (as for instance alla carlona 'carelessly', a perdifiato 'at the top of one's voice' registered under the headword carlona, perdifiato, words with combinatorial usage only. Italian corpora can help define the confines of the multilexical unit and establish possible variations, such as widespread elliptical uses. Coherent corpus-based decisions are in turn extremely valuable not only for lexicographers, but also for POS tagging of corpora in which the multilexical units are recognized and entered as a whole in addition to the single parts.

Monolingual lexicographies of Romance languages recognized and recorded - and record today - multiword units but do not promote them easily to headword status. Twentieth century Italian and Spanish monolingual lexicography, in particular, had very few multilexical units whereas French had a few more.

When Italian monolingual lexicography started, there was a different treatment of multilexical units. Here below are reproduced some Italian multilexical headwords from the Vocabolario degli Accademici della Crusca (1612). Academicians registered them as headwords with their own entry while in modern dictionaries they became part of the entries cavallo, tempo, ora, abbandono, se.[1]

> **A CAVALLO** posto avverbialm. Onde portare a cavallo. Lat. *vehere in equo*: e Stare a cavallo, o sopra un cavallo. Lat. *insidere equo*.
>
> **A TEMPO A TEMPO** avverbialm. Di quando in quando, di tempo in tempo, vicendevolmente. Lat. *statis temporibus*. ▸G. V. 8. 51. 2. Rimutandosi a tempo a tempo, con parte de' lor soldati.
>
> **DA ORA INNANZI**. avverbialm. per l' avvenire. Lat. *posthac in posterum*. ▸Petr. Son. 31. Da ora innanzi faticoso, e alto Luogo non sia.

[1] A cavallo 'on horseback', a tempo a tempo 'from time to time', da ora innanzi 'from now on' , da per se 'alone', in abbandono 'in a state of neglect; run-down', se non 'but, save', su stante 'up straight', tratto tratto 'at times'.

> **DA PER SE**. solo, senza compagnia. Lat. *perse, solus*. ▸G. V. 8. 87. 1. Prima s' andava ciascuna delle 21. arti da per se [cioè separatamente]
>
> **IN ABBANDONO**. Posto avverbialmente, vale senza cura, e senza custodia, ma s' accompagna co' verbi LASCIARE, e METTERE, che in lat. si direbbe *deserere*. ▸Bocc. n. 60. 9. Lasciata la camera, ec. in abbandono, là si calò.
>
> **SE NON**. Particella eccettuativa, e vale, fuorchè. Latin. *nisi praeter, quamquod*. ▸ Bocc. n. 43. 6. Che ne dobbiam noi far' altro, se non torgli quei panni, e quel ronzino, e impiccarlo?
>
> **SU STANTE**. Ritto in piede. ▸Nov. ant. 51. 4. Appresso il fece rizzare in su stante e gli cinse una bianca cintura.
>
> **TRATTO TRATTO**. Avverb. e vale di punto in punto, di momento in momento. ▸ Bocc. n. 81. 11. E parevagli tratto tratto, che Scannadio si dovesse levar ritto, e quivi scannar lui.

Such a ‘phrasal’ approach may be due to the fact that the Vocabolario degli Accademici della Crusca was a corpus-based dictionary as pointed out by Sabatini (2006:31): ‘ Ever since 1590 the Accademia della Crusca had been working on a comprehensive Dictionary of the Italian language, based on the study of texts mostly written in 14th century Florentine, but also including several later authors, not all Tuscan. The dictionary was published in Venice in 1612 and was the first modern European lexicographical undertaking in terms of its content and methods. The works of the authors quoted formed a balanced corpus. For each meaning a large context was supplied and there were frequent links to other related words and definitions. The Vocabolario acted as a centre of standardisation and identity of the language in Italy for centuries.’ Another lexicographic landmark the Dizionario della lingua italiana by N. Tommaseo and B. Bellini, concluded in 1879, was also rich in multilexical headwords, but again it was a big dictionary, mainly based on quotations from great Italian authors.

When and why the one-word headword policy prevailed is an interesting topic which deserves more space and attention. Here some hypothesis can be sketched.

The attitude changed with dictionaries in one volume for ‘everyday use’ such as Fanfani (1863) probably for space-saving reasons:

- the introduction of hundreds of technical and scientific terms which became headwords consumed space
- multilexical units were placed inside the body of the gloss of the first word of the multiword[2] because the dictionary was still meant to help the reader of ancient masterpieces but was mainly oriented towards the needs the needs of who had to write in Italian as a citizen of the recently unified state of Italy.

Also Barbera (2009) - discussing how Italian lexicography faced multiwords lemmatization - remarks that earlier in the process there might be a general change from a microstructure built on semantic grounds, i.e. to serve mainly understanding, towards a syntactically designed one more suitable for writing purposes. The change arrived first in small dictionaries for students:

[2] Or inside the body of the gloss of the word which was considered semantically more important and therefore likely to be the first searched by the user

multilexical units with adverbial or adjectival function were somehow awkward for lexicographers with prescriptive intentions.

In the mid-twentieth century the turn was so completed that even the *Grande dizionario della lingua italiana* (1961-2004) by Battaglia and Barberi Squarotti, a totally descriptive dictionary in 22 volumes, has almost no Italian multilexical headwords.[3]

In the first decade of the XXI century , as an evidence of the fact that only foreign multilexical units tend to be registered as multilexical headwords in Italian monolingual dictionaries, we can report that all the 25 *locuzioni*, i.e. multilexical headwords, registered in the last edition of Zingarelli (2009) as being entered in Italian language from 2001 onwards, are Anglo-American: from *advanced booking* (2002) to *You Tube* (2006)[4].

To date we find that in the most widespread monolingual Italian dictionaries, like Zingarelli or Sabatini-Coletti, the multiword *alla julienne* 'julienne' is listed under the headword **julienne**, because it is considered Italian. If it were registered in its totally French variant *à la julienne*, it would be registered as such, as a multilexical headword. And actually we find in Zingarelli **à la carte**, **à la coque** 'soft-boiled (egg)', **à la page** 'fashionable'. Sabatini-Coletti is consistent and lists **coque** and **carlona** as headwords since they are part of the Italian phrase *alla coque* e *alla carlona* and it states that they are used 'only in the phrase *alla c*.,.'. Zingarelli lists the headwords **à la coque** or **alla coque**, **all'erta** or **allerta** 'on the alert', **à jour** 'hemstitch', and sometimes has a third approach, that is, it lists **alla carlona** as a forward headword that refers to **carlona**, where one finds as part of speech label vc. (i.e. *voce* abbreviation for 'item') followed by the explanation 'only in the adverbial phrase ***alla carlona***'.

> **alla (1) o (poet.) a la** prep. art. f. sing. comp. di a (2) e la (1) Unita a un aggettivo femminile o a un sostantivo, forma numerose loc. avv. o agg. (con ellissi di 'modo', 'maniera', 'moda' ecc.): *all'antica, alla svelta, alla buona, alla garibaldina; alla fiorentina, alla milanese; alla brace, alla Bismark; alla panna*.
>
> **alla carlóna → carlona**
>
> **allacciamento (…)**
>
> **(…)**
>
> **carlona** [da *Carlo* Magno, rappresentato come un bonaccione nei poemi cavallereschi più tardi ☼ 1527] vc. ● Solo nella loc. avv. ***alla carlona***, alla buona, in fretta, con trascuratezza e in modo grossolano: *fare le cose alla carlona; tirar giù un lavoro alla carlona*. (Zingarelli 2009)

It is apparent from the above reported sample of macrostructure that Zingarelli, like all monolingual Italian dictionaries today, beyond the canonical form **a**, carries a headword for articulated prepositions and has the headword **alla** or (poetical) **a la** . In that lexicographic article it explains 'together with a feminine adjective or noun, forms many adjectival or adverbial phrases (but omits 'manner', 'way' etc.): *all'antica, alla svelta, alla buona, alla garibaldina; alla fiorentina, alla milanese; alla brace, alla Bismark; alla panna*.

[3] A cavallo is under cavallo, and also all the other examples from Vocabolario degli Accademici della Crusca are not multilexical headwords. Some are reduced to one-word headwords, as sustante, senon which forwards to sennò (which by the way is misleading because sennò has a different meaning, i.e. 'otherwise')

[4] The only exception is Tom Tom a registered trademark of a Dutch manufacturer which is used in Italian to refer to any type of satellite navigator.

A quick overview of monolingual Spanish dictionaries, from DRAE to the Diccionario VOX de Uso del español de America y España, and bilingual Spanish-English dictionaries reveals much the same situation as the Italian one: in the DRAE we find for instance *en diferido* ‘recorded broadcast’ ‘**diferido**.(Del part. de diferir).en ~ .1. loc. adj. Dicho de un programa de radio o de televisión: Que se emite con posterioridad a su grabación. U. t. c. loc. adv.’

Bilingual lexicography (particularly with English) is always more inclined to highlight multilexical units. The Harper Collins Spanish-English dictionary has a traditional printed entry and an electronic one which include *falsa alarma* ‘false alarm’ and the rest among examples and phraseology, while the Oxford Spanish Dictionary OSD in the electronic version carries lemmatization of all Spanish compounds, or at least of those which lead to noun phrases.

The lexicographical items concerning **falso** in monolingual Italian or Spanish dictionaries, for example, have a cross reference or phraseology, not sub-lemmas. Listed herewith is the Oxford Spanish Dictionary OSD entry: the printed version is on the right-hand side and on the left the word list as it appears in the computer version: that is with ‘sustantivos compuestos’ compounds promoted to multilexical headwords.

falso -sa *adj*
A1‹*billete*› counterfeit, forged; ‹*cuadro*› forged
2‹*documento*› (copiado) false, forged, fake; [omissis]…
B1 (no cierto) ‹*dato*/*nombre*/*declaración*› false;
eso es d, **nunca afirmé tal cosa** that [omissis]……
Compuestos
• **falsa alarma** *f* false alarm
• **falsa modestia** *f* false modesty
• **falso testimonio** *m* (Der) false testimony, perjury; **no levantar** dd(Relig) thou shalt not bear false witness

falso

falsa alarma femenino false alarm

falsa modestia femenino false modesty

falso testimonio masculino (Derecho) false testimony, perjury; **no levantar falso testimonio** (Religión) thou shalt not bear false witness

It might be thought that ostracism affects multilexical units which begin with a preposition, but a survey on adjectival phrases in the Zingarelli (2009) dictionary revealed 153 multilexical lemmas and sub-lemmas[5] which start with a preposition. Analyzing them we see that they have been listed as lemmas or sub-lemmas because they are borrowed from Latin, French or English. Italian adjectival phrases are few: the set of *non* + adjective, *non allineato,* ‘non-aligned’, *non vedente,* ‘visually handicapped’ and the more surprising *porta a porta,* ‘door-to-door’, *faccia a faccia* ‘face-to-face’ (so it reads that they are translations from English).

When the first element of a multilexical unit is not a preposition the approach is still the same: if it is a loan word then it is registered as a multilexical lemma, otherwise it is not and the multilexical unit goes inside another article.

In the Zingarelli 2009 edition, of the 795 *locuzioni nominali* ‘multilexical nouns’ reported, as many as 490 are English borrowings from *account executive* to *Yorkshire terrier*, 65 are French from *ancien régime* to *trait d’union,* ‘hyphen’, 29 Latin from *arbiter elegantiarum*

[5] Whereas in Sabatini-Coletti they appear to be fewer, 108, because many are sub-lemmas which are not identified by POS search ‘loc agg’.

‘arbiter of fashion’ to *schola cantorum, ‘*church choir’, 7 Spanish *(buen retiro, cante hondo, cha cha cha, cuba libre, el Niño, olla podrida, paso doble),* 1 German *Sturm und Drang,* 1 Russian *agit-prop* ‘political agitator’. The Italian multilexical units reported as multilexical headwords are onomatopeias (*glu glu,* ‘gurgle gurgle’, *gre gre* ‘croak croak’), lexicalized phrases (*cessate il fuoco* ‘ceasefire’, *chi va là,* ‘challenge’, *gratta e vinci* ‘scratch card’ or double verbs compound nouns *lecca lecca* ‘lollipop’, *mangia mangia,* ‘illicit gains’, *pigia pigia ‘*crush’). Lemmas like *gran premio,* ‘Grand Prix’, *mezzo punto,* ‘half-cross stitch’, *natura morta* ‘still life’ can be counted on one hand and what they have in common is that they are not formed by two nouns.

Multiwords formed by two nouns like *studio pilota*, ‘pilot study’, *nave scuola* ‘school ship’, *governo ombra* ‘shadow cabinet’ do not exist as lemmas in Italian dictionaries, because the second word *pilota, scuola* and *ombra* can be collocated with a limited number of nouns, so it becomes an ‘invariable postponed adjective’ in Zingarelli and Sabatini-Coletti. The Diccionario de Uso del español de America y España considers these uses of *piloto* appositions which do not agree in number with the noun they modify:’ **NOTA** Se construye en aposición a otro nombre con el que no concuerda en número: *piso piloto*; *experiencias piloto*; *este hospital será centro piloto de un experimento de diversificación energética*.’

Multilexical headwords which are not foreign borrowings nor nominal are rare in an Italian dictionary: it is striking to see how Italian monolingual lexicography treats phrasal verbs (for instance, *buttar giù,* ‘to dash off’, *buttar via,* ‘to throw away’, *metter su,* ‘to set up’, *fare fuori,* ‘to kill’, *venir meno* ‘to disappear’, etc.). They are not very many, approximately 250, however they have a high frequency use (see Cini 2008). Nonetheless they are listed within the printed entry, often mixed with the actual phraseology. The Sabatini-Coletti dictionary carries them as sub-lemmas (e.g. **buttare** has a section for *buttare addosso, buttare dentro, buttare fuori, buttare giù, buttare là* and *buttare via*). The most linguistically oriented of Italian monolingual dictionaries, i.e. De Mauro, goes so far as to use the POS *‘procomplementare’ i.e. procomplement verb* 131 times and lemmatizes forms like *correrceне, indovinarla, sfangarsela, starci, tornarsene, vedersela,* etc. . However it includes phrasal verbs like *buttare fuori, buttare giù, buttare là, buttare via* in the section *polirematiche* 'multiwords' of the lexicographic entry **buttare**, ‘to throw’, where they have a fair presence above all in the electronic version, but are mixed with light verb + noun collocations like *buttar sangue*, ‘to bleed’, and idioms like *buttare il bambino con l'acqua sporca* , ‘to throw the baby out with the bathwater’, or *buttare polvere negli occhi,* ‘to throw dust in sb.'s eyes’. They would also merit lemma status like the corresponding English verbs *throw about, throw away, throw back, throw in, throw out,* etc. found in monolingual and bilingual English dictionaries.

Buttare [omissis]
POLIREMATICHE:
buttare acqua sul fuoco: loc.v. **CO**
buttare all'aria: loc.v. **CO**
buttare al vento: loc.v. **CO**
buttare a mare: loc.v. **CO**
buttare a terra: loc.v. **CO**
buttare dalla finestra: loc.v. **CO**
buttare fango: loc.v. **CO**
buttare fuori: loc.v. **CO**
buttare giù: loc.v. **CO**
buttare il bambino con l'acqua sporca: loc.v. **CO**
buttare in faccia: loc.v. **CO**
buttare in, per aria: loc.v. **CO**
buttare in, sul volto: loc.v. **CO**
buttare là: loc.v. **CO**
buttare l'acqua sporca con il bambino dentro: loc.v. **CO**
buttare la tonaca: loc.v. **CO**
buttare le braccia al collo: loc.v. **CO**
buttare l'occhio: loc.v. **CO**
buttare olio sul fuoco: loc.v. **CO**
buttare polvere negli occhi: loc.v. **CO**
buttare sangue: loc.v. **CO**
buttare via: loc.v. **CO**
da buttare: loc.agg.inv. **CO**
(De Mauro 2000)

Lexicologists and applied linguists, above all those who work on corpora linguistics using computers, are against the process of hypostatization due to the fact that one sole word (usually the noun or adjective or verb) of a multilexical unit is worthy of becoming a lemma or sub-lemma in a dictionary.

Electronic dictionaries and the possibility of finding multilexical units in all the text, not just in the headwords field has taken the drama out of the issue of making a multilexical unit a (sub)headword, since in a lexical item on screen multilexical units go to a new line and thus are highlighted within an entry.

The matter of promoting multilexical units to (sub)headwords is still open to debate and opposition in monolingual paper dictionaries: the fact that users look up words in search windows on line with collocators which reduce polysemy or homonymy may convince Italian and Spanish lexicographers to introduce more multilexical headwords, rather than unnatural monolexical headwords.

Having approached the problem of where to list the multilexical unit in electronic dictionaries without solving it (because the unit is found quickly wherever it is treated), dictionaries, even with computer support, have to face the issue of entering a POS tag for the multilexical unit and so cannot be exempt from weighing the syntactical-semantic relations among the parts of the multilexical lemma.

One way out is avoiding it[6]: the Diccionario de Uso del español de America y España does not attribute a POS tag to *ex aequo,* it defines it as 'expresion latina'; instead it enters *dolce vita* as nombre femenino. The Sabatini–Coletti dictionary does not attribute a POS to 'artificial' headwords like *carlona* o *vanvera*, because it states in the explanatory note that they are found only in adverbial or adjectival phrases. The Zingarelli dictionary uses the 'non POS tag' *voce* (i.e. *entry)* 55 times in the 2010 edition, from *analda* to *vanvera,* most of the time because it did not want to list multilexical headwords which were easily POS tagged as adjectival, adverbial phrases etc. However, Zingarelli is still the dictionary which attempts most to face categorization by traditional POS.

Grammarians believe that studying word formation is an issue for morphologists and the latter investigate this phenomenon through corpora analysis. In the mean time lexicographers have to take hundreds of decisions every year, with or without linguists' support, and these decisions in turn become points of discussion for other linguists. Since a coherent theory is not always ascertainable behind dictionary decisions, but on the contrary criteria are debatable[7], it is high time that linguists and lexicographers (indeed linguists often act as lexicographers) decide to overcome the blank between the elements of the multilexical unit. Italian corpora can help define the confines of the multilexical unit and establish possible variations, in particular elliptical uses[8].

[6] If you reconsider the initial list of entries from Vocabolario degli Accademici della Crusca, you will notice that the Accademici did not always feel obliged to give a POS label to multilexical headwords'

[7] It is not declared in the forewords but it appears that the policy is to have a multilexical headword only when it is a foreign borrowing and preferably a nominal phrase.

[8] See Marello 1998 for a discussion of elliptical uses of nominal multilexical units, such as caccia 'literally hunting' for aereo da caccia 'fighter plane', Lettere for Facoltà di lettere 'Faculty of Arts', oro 'gold' for medaglia d'oro 'gold medal' and related changes of their morphology respectively in gender, number and countability.

Coherent corpus-based or corpus-driven decisions are in turn extremely valuable for POS tagging of corpora in which the multilexical units are recognized and entered as a whole in addition to the single parts[9]. A recent corpus the LIPSI (see Pandolfi 2009) has openly, with a separate list, declared the nature and number of multilexical units considered from the most frequent *per esempio* ‘for instance’, labelled as adverb, to *tale e quale* ‘exactly like’ considered as an adjective[10]. It is a first courageous step which marks perhaps a new beginning for multilexical units, towards a more realistic and user friendly lemmatization policy in Italian lexicography.

[9] In Corpus Taurinense multilexical units are recognized and entered as a whole in addition to the single parts. See www.corpora.unito.it. Barbera - C. Marello 2000 and Barbera (2009)

[10] We will not discuss here the POS annotation which was manual for multilexical units following the POStagset designed by CNR Pisa .

Bibliography

Dictionaries

Battaglia S., Bàrberi Squarotti G. (eds.) (1961-2004). *Grande dizionario della lingua italiana*. Torino: U.T.E.T., 21 voll + suppl.

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

Diccionario VOX de Uso del español de America y España en Cd-ROM

Diccionario de la Real Academia Española XXII edición en CD-ROM

Fanfani P. (1863). *Vocabolario dell'uso toscano.* Firenze: G. Barbera

Sabatini F., Coletti V. (2006). *Il Sabatini Coletti. Dizionario della lingua italiana.* Milano: Rizzoli-Larousse

Lo Zingarelli 2010 Vocabolario della lingua italiana di Nicola Zingarelli. Bologna: Zanichelli, 2009

Oxford Spanish Dictionary 2003 on CD third edition chief editors Carol Styles Carvajal Jane Horwood

Tommaseo N., Bellini B. (eds) (1865-1879). *Dizionario della lingua italiana*.Torino: Unione Tipografico-Editrice Torinese

Vocabolario degli Accademici della Crusca. (1612). Venezia: Giovanni Alberti

Other literature

Barbera M.- Marello C. (2000). 'Les lexies complexes et leur annotation morphosyntactique dans le Corpus Taurinense', intervento al convegno AFLA 2000, Paris, 6-8 luglio 2000, In *Révue française de linguistique appliquée*, vol. V-2 Décembre 2000 pp. 57-70.

Barbera M. (2009). *Schema e storia del 'corpus Taurinense'. Linguistica dei corpora dell'italiano antico.* Alessandria: Edizioni dell'Orso.

Cini M. (a cura di) (2008). *I verbi sintagmatici in italiano e nelle varietà dialettali. Stato dell'arte e prospettive di ricerca.* Frankfurt: Lang.

Grossmann M., Rainer F.(eds.) (2004). *La formazione delle parole in italiano*. Tübingen: Niemeyer

Pandolfi E.M. (2009). *LIPSI Lessico di frequenza dell'italiano parlato nella Svizzera italiana*. Bellinzona: Osservatorio Linguistico della Svizzera Italiana.

Sabatini F. (2006) 'La Storia dell'Italiano nella Prospettiva della Corpus Linguistics' In E. Corino', C. Marello, C. Onesti (eds.) *Atti del XII Congresso Internazionale di lessicografia, Torino, 6-9 settembre 2006 Proceedings XII Euralex International Congress*. Alessandria: Edizioni dell'Orso, I vol. 31-37.

Simone R. (1996). 'Esistono *verbi sintagmatici* in italiano?' In *Cuadernos de filologia italiana* 3, 47-61.

Marello C. (1998) 'What qualifies as an elliptical noun phrase in Italian. Opinions of grammarians, lexicographers and native speakers'. In J. Korzen, M. Herslund (eds.). *Clause Combining and Text Structure,* Copenhagen Studies in Language 22, Fredericksberg: Samfundlitteratur, 107-123.