The Lemmatisation of Lexically Variable Idioms: The Case of Italian-English Dictionaries

Chris Mulhall
Waterford Institute of Technology - University College Dublin

The choice of a suitable point of entry for an idiomatic expression is one of the most complex tasks a lexicographer faces throughout the compilation of a dictionary. This is further exacerbated by the possibility of lexical variation in certain expressions. This paper analyses twenty idioms with variable verbs (ten English / ten Italian) and twenty idioms with variable nouns (ten English / ten Italian) across six bilingual Italian-English dictionaries, Il Ragazzini *(ZIR) (2006),* Hoepli Grande Dizionario di Inglese *(HGDI) (2003),* Collins Sansoni Italian Dictionary *(CSID) (2003),* Oxford-Paravia Italian Dictionary *(OXID) (2001),* Il Sansoni Inglese *(ISI) (2006) and* Hazon Garzanti Inglese *(HGI) (2006). The analysis highlights a number of problems in the treatment of lexically variable idioms. Firstly, bilingual Italian-English dictionaries do not have a definitive approach to dealing with the problem of lexical variation. Secondly, the consistency and comprehensiveness in the coverage of lexical alternatives varies significantly both within and across the Italian-English and English-Italian sections of dictionaries. The totality of such differences suggests that a more systematic approach is required in order to achieve a greater consistency in the recording of the variable constituents of idioms.*

Introduction

According to Hartmann and James (1998: 83), lemmatisation is "a problem awaiting a comprehensive solution (attempted by computational approaches) in connection with the wider tasks such as how to choose a suitable headword from the constituents of a fixed expression." The term 'fixed expression' encompasses a subset of phrases referred to as idioms. The *Collins Cobuild Idioms Dictionary* (2004: v) defines an idiom as "a group of words which have a different meaning when used together from one it would if the meaning of each were taken individually." In addition to their unique semantic representation, a number of idioms can vary their lexical components. Gläser (1998:129) states that this occurs in phrases containing "constituents that allow variations within the constraints of the lexicological / phraseological system." A typical form of this is lexical variation. This occurs through the substitution of one lexical element with a synonymous equivalent, without affecting the overall meaning of the phrase. Hartmann and James (1998) highlight the difficulty in determining a suitable entry point for an idiom as a fixed unit, but this increases somewhat when lexical variation is possible in the expression.

The complex nature of this problem is particularly evident in the six bilingual Italian-English dictionaries used in the study. In general, the recording of fixed idioms follows a particularly haphazard arrangement in some dictionaries, possibly due to the absence of a lemmatisation strategy. A noteworthy feature of two dictionaries used in the study, the CSID (2003) and the ISI (2006), is that they clearly outline their lemmatisation strategy for phrasal units in the preface. The maxim states that phrases, idioms and proverbs in dictionaries are listed under the first key word in the expression (either verb, noun or adjective). In the case of phrasal units containing verbs of common usage and high phrasal frequency, such as *go*, *come*, *make*, *take,* etc in English and *andare*, *fare*, *prendere*, *venire,* etc in Italian, they are listed under the next key word in the phrase (noun or adjective).

Figures 1-4 illustrate the point(s) of entry for the selected idioms in each dictionary, verb, noun or adjective (V / N / ADJ) and the first, second or third alternatives if more than one verb, noun or adjective is possible (V1, V2, V3, N1, N2, N3). In some cases dictionaries use cross

references and citation forms, which are denoted by a superscript CR (^{CR}) and an italicised CIT (*CIT*) respectively. The recording of lexical variants under an invariable component is indicated by superscript V1, V2, N1, N2, etc (V^{N1N2}, N^{V1V2}, ADJ^{V1V2}, etc).

Variable Verb English	Il Ragazzini (2006)	Hoepli (2003)	Collins Sansoni (2003)	Oxford Paravia (2001)	Sansoni Inglese (2006)	Hazon Garzanti (2006)
To bend / stretch the rules	V1, V2, N^{V1V2}	N^{V1V2}, V2*CIT*	V1	V1, N^{V1}	V1	V1, N^{V1V2}
To blow / let off steam	V1, V2, N^{V1V2}	N^{V1V2}	V1,N^{V2}	V2, N^{V1V2}	V1,N^{V2}	V1, V2, N^{V2}
To cool / kick one's heels	V1, V2	N^{V1V2}	V1, V2	V2, N^{V1V2}	V1, V2	V1, V2
To drop / lower one's guard	V1, V2, N^{V1V2}	N^{V1V2}	N^{V1V2}	V2,N^{V1V2}	N^{V1V2}	N^{V2}
To fill / fit the bill	$V1^{CR\rightarrow V2}$, V2, N^{V1V2}	V2, N^{V1}	V1, V2	V2, N^{V1V2}	V1, V2	N^{V1V2}
To fling / throw down the gauntlet	V2, N^{V1V2}	N^{V2}	V1, V2, N^{V1V2}	V2, N^{V2}	V1, V2, N^{V1V2}	V2, N^{V1V2}
To hunt / search high and low	NOT LISTED	V1, V2, $ADJ1^{V2}$	V1, $ADJ1^{V1V2}$	V1, $ADJ1^{V1V2}$	V1, $ADJ1^{V1V2}$	V1, $ADJ1^{V1V2}$
To lift / raise the roof	N^{V2}	N^{V1V2}	V2	N^{V2}	V2	V2, N^{V1V2}
To pull / tear one's hair out	V2,N^{V2}	V2,N^{V2}	V2,N^{V1}	V2,N^{V2}	V2,N^{V1}	V2,N^{V2}
To raise / up the ante	N^{V1V2}	N^{V1V2}	N^{V2}	N^{V2}	N^{V2}	N^{V1}

Figure 1. The lemmatisation of the variable verb elements (English-Italian Sections)

Variable Verb Italian	Il Ragazzini (2006)	Hoepli (2003)	Collins Sansoni (2003)	Oxford Paravia (2001)	Sansoni Inglese (2006)	Hazon Garzanti (2006)
Ammazzare / Ingannare il tempo	V1, V2, N^{V1V2}	V1, V2	V1, V2	V1, V2, N^{V1V2}	V1, V2	V1, V2
Andare / Guardare per il sottile	$V1^{CR\rightarrow N}$, $V2^{CR\rightarrow N}$, N^{V1V2}	N^{V1V2}	V2, N^{V1}	V1, N^{V1CIT}	V2, N^{V1}	V1, V2, N^{V2}
Andare / Montare in bestia	V1, V2, N^{V1V2}	V2, N^{V1}	V2, N^{V1}	V2, N^{V1V2}	V2, N^{V1}	N^{V1}
Capire / Prendere fischi per fiaschi	V1,V2, $N1^{V2}$	V2, $N2^{V1}$	$N2^{V1}$	$N1^{V1}$, $N2^{V2}$	$N2^{V1}$	NOT LISTED
Crescere / Spuntare / Venire su come i funghi	N^{V1V2V3}	N^{V1V2}	V2, N^{V2V3}	V2,N^{V1}	V2, N^{V2V3}	N^{V1}
Dormire / Riposare sugli allori	V1, V2, N^{V2}	N^{V1V2}	V1, V2, N^{V2}	V1, V2, N^{V1V2}	V1, V2, N^{V2}	V2, N^{V2}

Drizzare / Raddrizzare le gambe ai cani	V2, N1^{V2,} N2^{V2}	V1, V2, N1^{V2}, N2^{V1}	V1, V2	V2, N1^{V2,} N2^{V2}	V1, V2	V2
Pungere / Toccare sul vivo	V1,V2, N^{V2}	V1, N^{V1V2}	V1,V2	V1, V2, N^{V1V2}	V1,V2	V1,V2, N^{V1V2}
Stringere / Tirare la cinghia	V2, N^{V2}	N^{V1V2}	V1,V2	V1,V2, N^{V1V2}	V1,V2	V1, N^{V1}
Stroncare / Troncare sul nascere	V1, V2, N^{V2}	V2, N^{V1}	V1, N^{V1}	V1, N^{V1}	V1, N^{V1}	N^{V1}

Figure 2. The lemmatisation of the variable verb elements (Italian-English Sections)

Lexical variation in the verb slot

The English-Italian and the Italian-English sections of the dictionaries show distinct differences in their recording of variable verb idioms. There are two possible lemmatisation strategies which achieve a complete coverage the variable elements, either entering the idiom under each variable constituent or alternatively recording all lexical alternatives under a fixed component. An ideal representation of this problem would be to record the idiom under all variable elements along with including its variations under the fixed component(s), but the practicalities of dictionary compilation and spatial constraints restrict this method being fully implemented in dictionaries.

The analysis of variable verbs reveals that no dictionary rigidly adopts either method. Furthermore, a contrast of the two approaches between the English-Italian and the Italian-English sections highlights a marked difference in their preferences of the available strategies and their overall coverage of variations. Firstly, the Italian-English sections in all six dictionaries list the variable verb elements as separate entries more frequently than the English-Italian section. Some examples of this are the ZIR (2006), which includes six variable verb Italian idioms under both variable verb entries, but does so with only four English idioms. Similarly, the CSID (2003) and ISI (2006) record variable verb idioms under all variants of five Italian expressions in comparison to only three English phrases. The most obvious disparity can be seen from a comparison of both sections in the OXID (2001), in which the Italian-English section enters four idioms (*ammazzare/ingannare il tempo*, *dormire/riposare sugli allori*, *pungere/toccare sul vivo* and *stringere/tirare la cinghia*) under both variable verbs, whereas its English-Italian section chooses one verb in the case of all ten English expressions under which to list the idiom.

In contrast, the English-Italian sections show a greater tendency to record the variable elements under the fixed component. This trend is evident in five of the six dictionaries with a higher number of English than Italian phrases recorded in this way. The only exception to this is the OXID (2001), which has five phrases in both sections listing all variable components under the fixed element. The HGI (2006) shows the greatest divergence with five English idioms (*to bend/stretch the rules*, *to fill/fit the bill*, *to fling/throw down the gauntlet*, *to hunt/search high and low* and *to lift/raise the roof*) in comparison to only one Italian idiom (*pungere/toccare sul vivo*) containing both variables under an invariable unit. Certain dictionaries give a complete representation of lexical variation by entering the idiom under all variants as well indicating the variations under the fixed components. The ZIR (2006) does this most comprehensively with 30% of the English idioms (*to bend/stretch the rules*, *to blow/let off steam* and *to drop/lower one's guard*) and 20% of Italian expressions (*ammazzare/ingannare il tempo* and *andare/montare in bestia*) all having a V1,V2,N^{V1V2} listing pattern. The OXID (2001) has the highest number of idioms with an ideal representation in the Italian-English section with 40% of its variable verb Italian idioms having a maximum coverage.

Variable Noun English	Il Ragazzini (2006)	Hoepli (2003)	Collins Sansoni (2003)	Oxford Paravia (2001)	Sansoni Inglese (2006)	Hazon Garzanti (2006)
To burn one's boats / bridges	V^{N1N2}, N1, N2	V^{N1N2}	V^{N1N2}	V^{N1N2}	V^{N1N2}	V^{N1N2}, N1, N2
To drag one's feet / heels	V^{N1}, N1	V^{N1N2}	V^{N1}	V^{N1N2}	V^{N1}	V^{N1}, N1
To give someone a dose / taste of their own medicine	$N3^{N1N2}$	$N3^{N1N2}$	N1, N2	N2, $N3^{N2}$	N1, N2	$N3^{N1N2}$
To hit the ceiling / roof	V^{N1N2}, N1, N2	V^{N1N2}	V^{N1N2}	V^{N1N2}, N1, N2	V^{N1N2}	V^{N1N2}, N1
To hit the hay / sack	V^{N1N2}, N1, N2	V^{N2}, N1	V^{N1N2}	N1, N2	V^{N1N2}	V^{N1N2}, N1, N2
To promise the earth / moon	V^{N1N2}, N2	V^{N1N2}	V^{N2}	V^{N1}	V^{N2}	V^{N1N2}, N1, N2
To reach for the moon / sky / stars	V^{N1}	V^{N3}, N2	V^{N3}	N3	V^{N3}	V^{N3}
To throw in the sponge / towel	V^{N1N2}, N1, N2	N1, N2	V^{N1N2}	V^{N1N2}, N2	V^{N1N2}	V^{N2}, N1, N2
To tip the balance / scales	V^{N1N2}, N1, N2	V^{N1N2}, N1*CIT*, N2	V^{N1N2}	V^{N1N2}	V^{N1N2}	V^{N1}, N2
To wear the pants / trousers	V^{N2}, N1, N2	V^{N1N2}	V^{N1N2}	N1,N2	V^{N1N2}	V^{N2}, N2

Figure 3. The lemmatisation of the variable noun elements (English-Italian Sections)

Variable Noun Italian	Il Ragazzini (2006)	Hoepli (2003)	Collins Sansoni (2003)	Oxford Paravia (2001)	Sansoni Inglese (2006)	Hazon Garzanti (2006)
Contare come il due a briscola / di picche	N1, N2	N1	V^{N1N2}, ADJ^{N1}	ADJ^{N1N2}, N2	V^{N1N2}, ADJ^{N1}	N1, N2
Dare corda / spago	N1, N2	N1, N2	N1, N2	N1, N2	N1, N2	N1, N2
Mettere cervello / giudizio	V^{N1N2}, N1, N2	N1,N2	N2	V^{N1N2},N2	N2	N1*CIT*, N2
Mettere la testa a partito / a posto	V^{N2N3}, $N1^{N2}$, N2	$N1^{N2N3}$, N2*CIT*, N3*CIT*	$N1^{N2N3}$, N2	$N1^{N3}$, N3	$N1^{N2N3}$, N2	V^{N2N3}, $N1^{N2}$, N2
Muovere mari e monti / cielo e terra	V^{N1N2}, N1, N3, N4	V^{N1N2}, N3	$V^{N1N2N3N4}$	V^{N3N4}, N3,N4	$V^{N1N2N3N4}$	N3, N4
Perdere la bussola / la tramontana	V^{N1}, N1, N2	N1, N2	V^{N1N2}, N2	V^{N1N2}, N1, N2	V^{N1N2}, N2	V^{N1N2}, N1, N2

Promettere la luna / mari e monti	V^{N2N3}, N1, N2, N3	V^{N2N3}, N1, N2*CIT*, N3	V^{N1N2N3}, N2	V^{N1N2N3}, N1, N3	V^{N1N2N3}, N2	V^{N2N3}, N1, N2, N3
Reggere / la candela / il lume / il moccolo	V^{N3}, N1, N2, N3	N1, N2, N3	V^{N1N2N3}, N3	V^{N1N3}, N2, N3	V^{N1N2N3}, N3	V^{N1N3}, N2, N3
Rendere pane per focaccia / la pariglia	V^{N1N2N3}, N1, N2, N3	N1, N3	V^{N1N2N3}	V^{N1N2N3}, N1, N2, N3	V^{N1N2N3}	N1, N2, N3
Serrare le file / i ranghi	V^{N1N2}, N2	N1, N2	V^{N1N2}	V^{N1N2}, N2	V^{N1N2}	V^{N1}, N2

Figure 4. The lemmatisation of the variable noun elements (Italian-English Sections)

Lexical variation in the noun slot

Generally, the recording of variable nouns gets a more extensive treatment in bilingual Italian-English dictionaries than verbs, but once again dictionaries differ in their depth and type of coverage the noun variants. Some dictionaries, such as the ZIR (2006), HGDI (2003) and HGI (2006), lemmatise more Italian idioms under all variable noun entries than English idioms. The other three dictionaries are more consistent in both languages; the CSID (2003) and the ISI (2006) listing one idiom in English and Italian under both nouns, whereas the OXID (2001) has entries under all noun variants for three Italian and three English idioms. The HGDI (2003) shows a particular imbalance in its coverage with seven Italian idioms (*dare corda/spago*, *mettere cervello/giudizio*, *mettere la testa a partito/posto*, *perdere la bussola/la tramontana*, *promettere la luna/mari e monti*, *reggere la candela/il lume/il moccolo* and *serrare le file/i ranghi*) and only two English idioms (*to throw in the towel/sponge* and *to tip the balance/scales*) inserted under each variable noun.

The extent to which dictionaries enter variable nouns under an invariable lemma in the both sections shows varying levels of consistency. In the English-Italian section all the noun variants of at least 50% of the selected phrases in each dictionary can be found under a fixed component, whereas it ranges from 10% to 80% in the Italian-English section. Dictionaries also vary in terms of their lemmatisation depending on the language in question. Evidence from the analysis reveals that the ZIR (2006), HGDI (2003) and HGI (2006) apply this method more frequently with English idioms, but the CSID (2003), OXID (2001) and ISI (2006) utilise this strategy on a greater scale with Italian idioms. In some dictionaries, major inconsistencies are obvious, for example, the HGDI (2003) enters all the variable nouns at the fixed point of seven English idioms (*to burn one's boats/bridges, to drag one's feet/heels, to give someone a dose/taste of their own medicine, to hit the ceiling/roof, to promise the earth/moon, to tip the scales/balance* and *to wear the pants/trousers*), but only one Italian idiom (*mettere la testa a partito/a posto*) receives the same treatment. Other dictionaries show greater consistency, such as the CSID (2003) and the ISI (2006), which have six English idioms and eight Italian idioms and the OXID (2001) with five English idioms and six Italian idioms following this method. Similar to the its variable verb coverage, the ZIR (2006) has the most ideal representation of variable nouns with 50% of the English idioms (*to burn one's boats/bridges, to hit the ceiling/roof, to hit the hay/sack, to tip the balance/scales* and *to wear the pants/trousers*) and 20% of Italian idioms (*mettere cervello/giudizio* and *rendere pan per focaccia/la pariglia*) having a complete coverage.

Conclusion

The wide diversity found in the approaches to the problem of lexical variation reaffirms the need for a systematic lemmatisation strategy. There are two particular trends that emerge from the analysis. Firstly, Italian-English sections show a greater tendency to enter the phrase under all the variable elements with 25 variable verb idioms and 25 variable noun idioms out of a combined total of 60 in each category following this method. This compares to 13/60 variable

verb idioms and 17/60 variable noun idioms with the same coverage in the English-Italian section. Secondly, the findings indicate that English-Italian sections show a preference for recording lexical variants under the fixed component, rather than under each element. Of the total of 60 in each category, 28 variable verb English idioms and 36 variable noun English idioms have all lexical alternatives listed at a fixed lemma in contrast to 14 variable verb Italian idioms and 29 variable noun Italian idioms. While this suggests a certain trend, it is not definitive, but does indicate the need for a lexicographical standard in recording lexical variations.

The two dictionaries with a definitive lemmatisation strategy, the CSID (2003) and the ISI (2006) contain the most extensive coverage of listing variable noun elements at a fixed point with an average of 70% of English and Italian idioms in this category following this pattern. In contrast, the listing under separate entries shows particular weakness with only 10% of idioms in both sections having entries under all variable nouns. In terms of their coverage of verb variants, 30% of variable verb English idioms and 50% of variable verb Italian idioms have listings under each variable. Conversely, the recording of verbs variants under a fixed component was lower with 30% of English idioms and no Italian idiom having all verb variants inserted under a fixed component. This may suggest that when designing a lemmatisation strategy the integration of a mechanism that allows the flexibility to fully encompass lexical variation may offer a potential solution to addressing this longstanding lexicographical problem.

References

Gläser, R. (1998). "The Stylistic Potential of Phraseological Units in the Light of Genre Analysis". In Cowie, A. P. (ed.). *Phraseology Theory*, *Analysis*, *and Applications.* Oxford: Oxford University Press. 125-143.

Hartmann, R. R. K.; James, G. (1998). *Dictionary of Lexicography*. London: Routledge.

Dictionaries

Collins COBUILD Idioms Dictionary. Glasgow: HarperCollins Publishers, 2004.

Collins-Sansoni Italian Dictionary. 4th ed. Milano: Edigeo, 2003.

Garzanti Hazon Dizionario Inglese. Milano: Garzanti Linguistica, 2005.

Grande Dizionario di Inglese. 2nd ed. Milano: Ulrico Hoepli Editore, 2002.

Il Ragazzini. 4th ed. Bologna: Zanichelli, 2005.

Il Sansoni Inglese. 5th ed. Milano: Edigeo, 2006.

Oxford-Paravia Italian Dictionary. Trento: Paravia Bruno Mondadori Editori and Oxford University Press, 2001.