# Compiling bilingual dictionaries: AI-Assisted translation of Italian Multiword Expressions into English and French

**Annalisa Greco[1], Matteo Delsanto[2], Andrea Di Fabio[2], Lorenzo Mori[2], Cristina Onesti[1], Daniele Paolo Radicioni[2], Calogero Jerik Scozzaro[2]**

[1] Università degli Studi di Torino, Dipartimento di Lingue e Letterature straniere e Culture Moderne, via Sant'Ottavio, 18, 10124, Torino

[2] Università degli Studi di Torino, Dipartimento di Informatica, Corso Svizzera, 185, 10149, Torino

E-mails: annalisa.greco@unito.it, matteo.delsanto@unito.it, andrea.difabio@unito.it, lorenzo.mori31@edu.unito.it, cristina.onesti@unito.it, daniele.radicioni@unito.it, calogerojerik.scozzaro@unito.it

## Abstract

The present research explores the use of large language models (LLMs) in digital lexicography, specifically for translating Italian multiword expressions (MWEs) into English and French. The study aims to assess the capability of contemporary LLMs in providing accurate and reliable translation equivalents, examples and definitions of Italian MWEs into English and French, while also evaluating the need for expert validation in refining AI-generated lexicographic resources. We seek to develop a digital resource tailored for language learners, offering frequently attested translations.

Methodologically, 120 expressions were evaluated by human experts and compared across two LLMs (Gemini 2.0 Flash and Mistral-Large-2411) using different metrics aimed at assessing including correctness, accuracy and contextual suitability, along with the capacity to produce meaning explanations and usage examples. Results show that English translations received higher expert ratings than French ones, with high correlation between human and AI evaluations in the case of English, and significantly lower agreement in the case of French translations. The findings indicate that LLMs provide generally reliable translations, though expert oversight remains crucial.

**Keywords:** multiword expressions; large language models; AI-assisted translation; bilingual dictionaries; dictionary writing system/dictionary-making process

## 1. Introduction

In the field of digital lexicography, the integration of artificial intelligence and natural language processing tools offers new opportunities to enhance the process of compiling dictionaries and to improve user experience. The ability of large language models (LLMs) to process and generate linguistic data at scale presents both advantages and challenges in lexicographic applications. Recent trends in contemporary lexicography show that the discipline is shifting towards an interdisciplinary approach also accounting for natural language processing techniques (Lew, 2024), and dealing with elaboration tasks, such as automatic summarization, translation, and question

answering.

In this study we single out the topic of translating MWEs such as idioms, collocations, and fixed expressions which pose significant challenges in translation because their meaning is frequently non-compositional, that is not directly inferable from their constituent words. LLMs play an increasingly important role in translating MWEs thanks to their capacity to process and generate text with contextual awareness (whereby meaning is inferred from the context rather than relying solely on word-level alignments), semantic flexibility, and cross-linguistic generalization. Their architectural design, particularly the Transformer architecture, allows LLMs to attend to multiple parts of an input sentence simultaneously using self-attention mechanisms: on these bases LLMs grasp the semantic and syntactic context necessary for correctly interpreting and translating MWEs. Additionally, the pre-training carried out on massive multilingual corpora equips LLMs with deep cross-linguistic representations, while fine-tuning is customarily adopted to further refine these capabilities for specific tasks or domains.

To fully exploit the generative capacities of LLMs it is preliminarily needed to assess their strengths and limitations: our research questions can thus be cast as follows:

- how accurately do LLMs like Gemini and Mistral generate translations, examples, and definitions, that are the main elements needed to craft an online dictionary?
- in how far does LLMs' judgement on the quality of a text correlate with human judgement?

We show that while high-quality translations can be achieved using two popular models, the performance on different types of MWEs varies substantially across models and languages. These findings therefore highlight how the models can be considered extremely useful for the creation of AI-assisted[1] multilingual resources, while still requiring supervision by human experts.

The long-standing aim is at creating a digital tool for language learners that offers frequently used, contextually appropriate, and expert-validated translations of complex MWEs, helping to bridge the gap between existing dictionaries and next-generation AI-assisted tools.

After surveying the closest approaches and research efforts available in literature, we briefly introduce the theoretical framework underlying the MWEs taken into account, and illustrate the materials employed for the evaluation, describe the experimental design, provide the obtained results and discuss them. We conclude by mentioning the

[1] Provided that by "AI-generated text" we refer to linguistic output produced by generative models, the focus of this work is instead on the notion of "AI-assisted". In this case, the texts generated by LLMs are reviewed by human experts to ensure a critical assessment of the models' linguistic output.

limitations of the present work and outlining the future work.

## 2. Related work

MWEs are linguistic objects consisting of two or more words that act "as a single unit at some level of linguistic analysis" (Calzolari *et al.*, 2002), showing an idiosyncratic nature in comparison to free word combinations. Due to their fixed or semifixed nature, these linguistic phenomena raise a number of challenges (see the recent overview in Giouli & Barbu Mititelu, 2024; for the Italian language Faloppa, 2011; Marello, 1996; Voghera, 2004). However, MWEs are pervasive and very frequent in everyday language; hence, they are highly useful for language learners and continue to represent a challenging area of research in bilingual lexicography. In a MWE, "the syntactic or semantic properties of the whole expression cannot be derived from its parts" (Villavicencio *et al.*, 2015). Such non-compositionality means unpredictability, accompanied by the non-substitutability of synonym words and a general lack of homogeneity. Some expressions do not present internal variation, while other ones allow various degrees of internal modification (e.g., the English expression *spill the beans* allows occurrences like: *spill/spilt the/some of/several/all the beans*).

This heterogeneous class of phenomena has implied a certain definitional ambiguity, which constitutes a further difficulty for a clear analysis of MWEs, therefore affecting their effective lexicographic treatment. According to their categorical properties and their degrees of fixedness, various types of MWEs have been classified: 'collocations', 'fixed expressions', 'phrasal verbs', 'idiomatic expressions'*, 'polirematiche'* in the Italian tradition, etc. (see, among others, Gantar *et al.*, 2019). However, as a consequence, clear boundaries between different categories are often lost, adopting 'MWE' as an umbrella term that encompasses a significantly varied set of items (Masini, 2019).

MWEs are often recognized as posing significant challenges for successful computational treatment as well (Sag *et al.*, 2002*;* Villavicencio *et al.*, 2015). In particular, no word-for-word translation is possible for them, as their non-compositional nature requires models to adopt strategies for accurately identifying their boundaries and capturing their intended meaning. Given the importance of MWEs not only in lexicography[2], but also in tasks such as machine translation, word sense disambiguation, and information retrieval, there is a clear and growing interest in developing AI-assisted solutions capable of effectively handling such complex linguistic phenomena.

Several systems and approaches have been presented in the literature, aimed at the automatic construction of resources to handle MWEs[3]. The paper by Garcia *et al.* (2019) describes the creation of an automatically built multilingual online dictionary

[2] See, for example, the interesting work of Orenha-Ottaiano (2017) regarding the construction of a bilingual dictionary of collocations.

[3] 'MWE' is employed here as an umbrella term covering other definitions such as 'collocations' (as in Garcia *et al.*, 2019; Orenha-Ottaiano, 2017).

of collocations in English, Portuguese, and Spanish. It includes verb-object, adjective-noun, and noun-noun combinations of words. Using dependency parsing, statistical measures, and unsupervised cross-lingual semantic models, collocations are extracted from large corpora and aligned across languages. The dictionary offers ranked translation equivalents and also serves as a monolingual resource for native speakers and learners of foreign languages. The work by Orenha-Ottaiano *et al.* (2021) outlines the initial development stages of PLATCOL, an online platform for multilingual collocations dictionaries tailored to different users' needs. Covering verbal, adjectival, nominal, and adverbial collocations, the methodology combines automatic extraction using NLP tools across five languages with expert post-editing. The approach also relies on statistical measures, distributional semantics to organize collocations by sense and to provide corpus-based examples.

Generative models have been investigated in many ways, to assess their efficacy in learning languages, to assess the translation quality, and as tools to support corpus linguistics. The study by Lew *et al.* (2024) explores whether LLM-based chatbots like ChatGPT can outperform traditional dictionaries as lexical tools for language learners. It compares ChatGPT with the monolingual Longman Dictionary and the bilingual Diki.pl in helping 166 university students (B2-C1 level) understand and use 40 uncommon English phrasal verbs. Results show that ChatGPT is more effective than both dictionaries for language production, and better than the monolingual dictionary, but not the bilingual one, for comprehension. The research illustrated in Mohammed (2025) compares translation quality between Arabic and English using rule-based systems, neural machine translation tools (e.g., Google Translate), and large language models (e.g., ChatGPT). It evaluates performance through both quantitative metrics (BLEU, TER, chrF) and qualitative analysis (House's model), focusing on idiomatic, colloquial, and technical language. Results show that while automatic metrics are helpful, they often miss semantic and contextual subtleties. ChatGPT performs better in handling nuance, but all systems still require human post-editing. A hybrid human-AI approach is thus recommended, especially for complex, culturally rich languages like Arabic. The work by Uchida (2024) explores how early LLMs, specifically ChatGPT 3.5, can support corpus linguistics. Considered tasks include generating word frequency lists, collocations, grammatical patterns, and identifying genres, with outputs compared to COCA corpus data[4]. While genre identification was weak, LLM outputs aligned fairly well with COCA on frequency lists, collocations, and grammatical patterns. Even mismatched items tended to be high-frequency. The Author concludes that although not yet suitable for rigorous research, LLMs are a convenient tool for identifying general trends and supporting learners, particularly due to their ability to search at the phrase level.

[4] https://www.english-corpora.org/coca/

## 3. AI-assisted MWEs translation

Our methodological approach includes a sample-based evaluation of LLM-generated translations through both human assessment and comparative analysis using models. The human evaluation involved domain experts rating the accuracy and appropriateness of LLM output, assessing Correctness, Accuracy and Functional Equivalence, Class Adequacy, Context Usage Suitability, Completeness or Superabundance, Enhancement (see 4.2.1). We gave a first evaluation of three different types of expressions considered by selecting overall 120 MWEs. The types of MWEs of interest are illustrated in the next Section (see 3.1).

### 3.1 Revisited MWEs' categories

Greco (in prep.) has recently proposed a classification of three different sorts of Italian MWEs, in order to unify the various definitions of MWEs existing in literature.

The multiword expressions we used for our evaluations are drawn from two paper specialised dictionaries and one online dictionary of Italian: *Tiberii* (2012), *Russo* (2010), *De Mauro-Internazionale*[5] (respectively specialised in Italian *collocazioni*, *modi di dire* and *polirematiche*[6]). When comparing different dictionaries, some MWEs may appear in more than one specialised dictionary, having therefore been labelled differently (e.g., *acqua e sapone* that in English could be translated as 'natural beauty', is recorded in all the three dictionaries considered here). Although specialised, such dictionaries label expressions according to different parameters and this inconsistency can be the source of perplexity and debate in different fields, such as lexicology (what is the boundary between one expression and another?); lexicography (what label to give to expressions recorded in a dictionary?); translation (how do I translate the expression? Literally? And, if not, how do I find the exact equivalent expressions in a target language?); glottodidactics - or language teaching (phraseological competence and metalinguistic awareness are two relevant factors for learning a foreign language); and, last but not least, natural language processing and in the use of tools such as LLMs for the compilation of a digital dictionary for learners of Italian as a Second and/or Foreign Language (L2/FL).

For this reason, we aimed to simplify definitional issues with a twofold objective: (a)

[5] Use the following link to reach the Italian online dictionary De Mauro-Internazionale: https://dizionario.internazionale.it

[6] We do not translate names in English, nor in French, because they do not always correspond in the two other languages. For instance, 'collocation' in English includes 'idioms' and 'sayings'; while in Italian 'collocazione' does not. The term '(espressioni) polirematiche' was specifically introduced by Tullio De Mauro (1999), who defines it as 'A group of words that has a unified meaning, not deducible from the meanings of the individual words that compose it, both in everyday usage and in technical-specialist languages'. As Hjelmslev states, in languages, reflections of the imagery net on matter are not homogeneous, and consequently the words are not identical (see Hjelmslev, 1968:62). We can say the same for our names too.

to find a lexicographic criterion for marking expressions in online bilingual dictionaries; (b) to find a criterion for understanding the quality of MWEs' translations proposed by LLMs. We report here on our proposal for new definitions of subcategories of MWEs, needed for Italian specific features (differently from other perspective, e.g. Burger 2010 concerning German). Given the use of "multiword expression" as an umbrella term, referring to different kinds of 'combinations of words' (see 2.1), we decided to adhere to the two overmentioned objectives and give prominence to one of the salient features of MWEs, namely their degree of metaphoricity - leaving aside other issues that are nevertheless useful to examine all these categories (such as collocability, idiomaticity, compositionality, figurative meaning, fixedness, in Gantar *et al.*, 2019:140; for Italian, see also Konecny, 2010). In short, we relied on a criterion of practicality: make these expressions easy to classify in a digital dictionary - and thus make them recognisable by learners of Italian, and also, help to evaluate translations provided by Gemini and Mistral.

We have therefore identified seven degrees of metaphoricity and, in doing so, defined seven new (sub)categories of Italian MWEs. This seemed to us the only, or perhaps the first, criterion to refer to for a classification of MWEs that could be useful to scholars of fields such as lexicography, NLP, glottodidactics, translation, a classification criterion that simplifies interdisciplinary work and allows a simple, intuitive, and useful way of defining these categories of words. Nevertheless, we report here just three new categories, the ones useful to describe our study, corresponding to the 2nd, the 3rd and 5th levels in our metaphorical scale.

*Co-occorrenze semplici* (=*Literal MWEs* - literal meaning).
Definition: these expressions have a literal meaning; every single word can be combined with other words. The metaphor is absent.
Example1: *agenzia di viaggi* > EN 'travel agency'.
Example2: *alzare la mano* > EN 'raise your hand'.
Explanation 1&2: in both examples, the two expressions retain a literal meaning.

*Co-occorrenze figurate* (=*Figurative MWEs* - figurative meaning).
Definition: the meaning of these expressions is metaphorical and goes beyond the literal one; it is derived from the result of a set of words; alternatively, in these expressions, at least one of the words is used in a metaphorical sense.
Example1: (*ragazza) acqua e sapone* (=someone, often a girl, who doesn't need any makeup to appear naturally beautiful).
Explanation1: in this first example, the combination of words *ragazza* plus *acqua* and *sapone* (respectively, in English, 'girl', 'water', 'soap') give rise to a third metaphorical meaning: a girl who is naturally beautiful.
Example2: *tenere aggiornato* > EN 'to keep someone informed about something'.
Explanation2: the first meaning of *tenere* (EN 'keep'), in Italian, is literally 'to have in the hand or between the hands'. As, in this example, you don't 'keep' (=*tenere*) anything literally between hands - but just in a metaphorical sense, you keep informed

your interlocutor about some news - here *tenere* acquires a metaphorical sense. So, *tenere* plus *aggiornato* get a new meaning: 'to keep someone informed about something'.

*Co-occorrenze idiomatiche immediate* (=*Immediate idiomatic MWEs* - idiomatic meaning).
Definition: their meaning is expressed using metaphors making a comparison between the speaker's reality and another reality where the image evoked takes place. Some of these expressions contain the Italian adverb of manner *come* (*libero come l'aria* > EN 'like'/'as': e.g. *(as) free as a bird*); otherwise, they imply it by ellipsis as in the example *avere il cuore d'oro* > EN 'to have a heart of gold'.
Example1: *libero come l'aria* > EN '(as) free as a bird'.
Explanation1: in this first example, the speaker makes a comparison between his/her interlocutor and *l'aria* (EN 'the air' - corresponding to 'a bird' in the English expression); the objective is to point out their shared quality of freedom.
Example2: *avere il cuore d'oro* > EN 'to have a heart of gold'.
Explanation2: in this example, the speaker makes a comparison between the heart, symbol of love, empathy, and so on, and the gold (=richness, abundance). So, a heart made of gold represents the quality of generosity.
From the 1st to 7th level[7] there is a gradual move away from the literal meaning of the expression and from proximity to the contingent reality towards other realities[8]. In the following Section we will illustrate the creation of the dataset, and particularly, the selection of expressions; their evaluation; and the usefulness of these new categories for our study.

## 4. Evaluation

### 4.1 Materials

We first selected a list of lemmas extracted from the 'Lexical Profile' for Italian as for the CEFR (Common European Framework of Reference for Languages), and namely from the B2 knowledge level list of words, and from letters A to C[9] (for headwords and translations of all Italian MWEs in French and English, see Annex I).

Then, we considered MWEs where at least one word was included in this list (taken from three specialised dictionaries of Italian, as illustrated in 3.1). So, we extracted 3300 MWEs and asked Gemini and Mistral for their translations, explanations and examples into French and English. We selected[10] from the two lists: 40 *collocazioni*;

[7] For a Table of Categories, see Annex III; for further information on new categories, see Greco, in preparation).

[8] In language philosophy, a metaphor typically establishes a connection between two realms of experience that are away from us. It establishes links that are not immediately and so not easily comprehensible in common language.

[9] We selected full words and mostly nouns - just one adjective: *bello*; and one verb: *dare*. Three extra headwords beginning with other letters: *dare*, *oro*, *stampa*.

[10] We aim to conduct more extensive work in the future: the other MWEs could be evaluated

40 *espressioni polirematiche*; and 40 *modi di dire* (which will fall, in this study, respectively in our new Literal, Figurative and Immediate idiomatic MWEs categories), in order to analyze a representative sample of the phenomenon in the Italian language and assess their practical implications for language learners.

### 4.2 Experimental Design and Procedure

Two models freely accessible via API were identified: on these bases, we selected Gemini (Gemini 2.0 Flash) and Mistral (Mistral-Large-2411). Each model was queried by using the prompts described below for each of the designed tasks, which are also detailed in the following sections.

We began by pre-processing and cleaning the source data, resolving issues such as MWEs appearing on the same line and repeated or missing items. From this cleaned dataset, we selected 40 expressions across three categories (*collocazioni*, *espressioni polirematiche*, *modi di dire*) in two languages (French and English), for evaluation with two large language models (Gemini and Mistral). The extraction and evaluation were carried out separately for each language and category

Then we rated translations, definitions and usage examples.

#### 4.2.1 Evaluation criteria

The output produced by the models includes *i*) the translation of the input expression, *ii*) the explanation of the meaning of the expression, and *iii*) an example of usage.

Translations were rated along six different *axes*: Correctness, Accuracy and Functional Equivalence, Class Adequacy, Context Usage Suitability, Completeness or Superabundance, and Enhancement. Both the explanation and the example usage received a concise rating estimating the correctness and appropriateness of such linguistic productions.

Such criteria were selected drawing inspiration from the metrics used in the field of translation. However, the classic parameters used to evaluate the quality of a translation are not entirely useful because, in our case, the comparative processes are applied to a much smaller portion of text (for example, here the notion of readability is not useful). Moreover, MWEs have more specific implications (for instance, equivalence might instead be useful in evaluating the communicative intention). It was therefore considered necessary to create *ad hoc* metrics with specifically lexicographic purposes, which are illustrated in the following:

- Correctness: is the rendering in the target language clear and correct in terms

later.

of spelling, morphology and syntax?

- Accuracy and Functional Equivalence: is the message from the source language to the target language the same? Are meaning, communicative intention and connotation maintained?
- Class Adequacy: does the translation rendering belong to the same class of word combinations? Is there an adaptation in gradation of meaning?[11]
- Context Usage Suitability: is the usage identical in both languages?[12]
- Completeness or Superabundance: is the translation rendering complete? Is there additional information or something missing?
- Enhancement: does the rendering contain elements of refinement of meaning, such as punctuation marks, capitalization, etc.?[13]

#### 4.2.2 Task 1: human assessment

Two expert lexicographers rated the output generated through the models at stake, one addressing English translations, explanations, and usage examples, and the other evaluating the French ones. They initially defined and agreed upon criteria for evaluating the translations and other outputs produced by the models under consideration. They used as a reference the translations provided by one of the following digital dictionaries: for English, The Collins Online Dictionary, the Merriam-Webster (monolingual dictionaries); Oxford Learner's Dictionaries (learner's dictionaries); for French *Trésor de la langue française* (TLFi); *Larousse*; *Le Grand Robert*; *Le Petit Robert*; *Le Robert Dico en ligne* (monolingual dictionaries). If none of the proposed dictionaries provided an equivalent, they then searched on the web and included the expression that seemed most appropriate, flagging it as featured by lack of attestation in lexicographic resources.

After completing the evaluation, they discussed the scores to harmonize their assessments.

Translations, explanations and usage examples were rated based on a scale ranging

[11] The *Class Adequacy* parameter is not among those normally proposed by scholars, but it is useful for providing additional information regarding both category adaptation and grading adaptation. Category adaptation indicates whether the expression is translated as a combination of words belonging to the same category as the original (e.g., *lupus in fabula* is not the same as saying "speak of the devil and he appears" or "we were just talking about you and here you are!"). Grading adaptation refers to expressing the same thing with a different degree of intensity or meaning (e.g., saying "reluctantly" is not the same as saying "with a heavy heart").

[12] This parameter does not include only style and tone, but also other variations based on social or communicative factors (e.g., register) - elements that we often find in the most authoritative dictionaries.

[13] It should be stressed that not in all cases enhancement applies, in that a given expression may not require any form of enhancement: in these cases, raters had simply to annotate that enhancement was unnecessary (*unnec.* value). This is the case for most translations (e.g. *canone di abbonamento* > EN 'subscription fee', FR 'frais d'abonnement'). Enhancement instead needs to be pointed out in a few cases (*accordo di riservatezza* > EN 'non-disclosure agreement', FR 'accord de confidentialité').

over the interval [-2, 2]. Additionally, for the Enhancement criterion, we added a sixth possible value, *unnec.*, reporting that for the translation at hand enhancement was not needed.

#### 4.2.3 Task 2: cross-model assessment

In the second phase of the experiment, we took the output generated by each of the models and asked the other model to evaluate it, based on the instructions provided in the prompt (see Annex II).

The models were requested to provide their assessment through a Likert scale from 1 to 5; this approach aligns with a common practice in the literature (Lee *et al.*, 2025). Both sets of scores were then rescaled to the [0, 1] range. In this setting, the models were fed with the expected translation (so to force them to grade the distance between two translations for the same MWE), both when queried to assess the translation, definition, and usage example.

### 4.3 Results

The results of the human assessment task, provided in Table 1, show that the translations produced by the LLMs received overall positive assessment by expert lexicographers. First of all, we observe that both models were acknowledged to provide better translations for the English language than for French: the average values recorded for all but Enhancement criteria (third-to-last column)[14] drop from 0.91 (EN) to 0.81 (FR) for Gemini, and from 0.86 (EN) to 0.80 (FR) for Mistral. A more detailed analysis of the translation shows that criterion 4, i.e. Context Usage Suitability, is in line with the other scores in the English translations, whilst it significantly drops in French translations; this effect was observed in the translations by both models.

If we consider the table rows, both models seem to suffer from analogous limitations: when dealing with English, the definitions (for *co-occorrenze semplici*, *co-occorrenze figurate*, *co-occorrenze idiomatiche immediate*) obtain ratings that are substantially on a par with those for the translations, with examples seeming less reliable (larger drops are associated to ratings for the items from the *co-occorrenze figurate*). When looking at the results for French, both models score significantly lower in the generation of definitions, as well as in the creation of usage examples.

[14] The average values computed in this column do not include the Enhancement criterion, since this was not analyzed in all the considered linguistic samples; therefore, these values are to some extent less general than the first five.

**GEMINI**

| | | TRANSLATION | | | | | | | DEFINITION | EXAMPLES |
|---|---|---|---|---|---|---|---|---|---|---|
| | METRIC | 1 | 2 | 3 | 4 | 5 | 6 | AVG(1,5) | | |
| Co-occ. semplici (EN) | AVG | 0.98 | 0.92 | 0.98 | 0.94 | 0.92 | 0.79 | 0.95 | 0.96 | 0.86 |
| | STDEV | 0.11 | 0.19 | 0.10 | 0.19 | 0.21 | 0.14 | 0.16 | 0.15 | 0.32 |
| Co-occ. figurate (EN) | AVG | 0.95 | 0.85 | 0.91 | 0.89 | 0.86 | 0.64 | 0.89 | 0.86 | 0.81 |
| | STDEV | 0.22 | 0.33 | 0.26 | 0.28 | 0.30 | 0.20 | 0.28 | 0.32 | 0.36 |
| Co-occ. idiomatiche immediate (EN) | AVG | 0.99 | 0.80 | 0.88 | 0.86 | 0.93 | 0.56 | 0.89 | 0.86 | 0.84 |
| | STDEV | 0.04 | 0.34 | 0.27 | 0.31 | 0.20 | 0.43 | 0.23 | 0.30 | 0.32 |
| **AVERAGE** | AVG | 0.97 | 0.86 | 0.92 | 0.90 | 0.90 | 0.66 | 0.91 | 0.89 | 0.83 |
| | STDEV | 0.12 | 0.29 | 0.21 | 0.26 | 0.24 | 0.25 | 0.22 | 0.26 | 0.34 |
| Co-occ. semplici (FR) | AVG | 0.89 | 0.87 | 0.86 | 0.72 | 0.88 | 0.88 | 0.85 | 0.69 | 0.56 |
| | STDEV | 0.23 | 0.23 | 0.30 | 0.23 | 0.23 | 0.18 | 0.25 | 0.21 | 0.31 |
| Co-occ. figurate (FR) | AVG | 0.86 | 0.86 | 0.89 | 0.66 | 0.86 | 0.75 | 0.82 | 0.69 | 0.62 |
| | STDEV | 0.29 | 0.28 | 0.26 | 0.31 | 0.27 | 0.00 | 0.28 | 0.26 | 0.32 |
| Co-occ. idiomatiche immediate (FR) | AVG | 0.82 | 0.78 | 0.72 | 0.68 | 0.79 | 0.95 | 0.76 | 0.60 | 0.64 |
| | STDEV | 0.29 | 0.31 | 0.42 | 0.28 | 0.34 | 0.11 | 0.33 | 0.60 | 0.64 |
| **AVERAGE** | AVG | 0.86 | 0.83 | 0.82 | 0.69 | 0.84 | 0.86 | 0.81 | 0.66 | 0.61 |
| | STDEV | 0.27 | 0.27 | 0.33 | 0.28 | 0.28 | 0.10 | 0.29 | 0.36 | 0.42 |

**MISTRAL**

| | | TRANSLATION | | | | | | | DEFINITION | EXAMPLES |
|---|---|---|---|---|---|---|---|---|---|---|
| | METRIC | 1 | 2 | 3 | 4 | 5 | 6 | AVG(1,5) | | |
| Co-occ. semplici (EN) | AVG | 0.96 | 0.78 | 0.96 | 0.89 | 0.87 | 0.75 | 0.89 | 0.88 | 0.84 |
| | STDEV | 0.14 | 0.35 | 0.11 | 0.25 | 0.29 | 0.12 | 0.23 | 0.29 | 0.33 |
| Co-occ. figurate (EN) | AVG | 0.91 | 0.74 | 0.86 | 0.64 | 0.80 | 0.00 | 0.79 | 0.82 | 0.71 |
| | STDEV | 0.25 | 0.40 | 0.29 | 0.47 | 0.37 | 0.31 | 0.36 | 0.38 | 0.43 |
| Co-occ. idiomatiche immediate (EN) | AVG | 1.00 | 0.78 | 0.87 | 0.88 | 0.97 | 0.50 | 0.90 | 0.87 | 0.84 |
| | STDEV | 0.00 | 0.37 | 0.28 | 0.29 | 0.12 | 0.00 | 0.21 | 0.30 | 0.36 |
| **AVERAGE** | AVG | 0.96 | 0.77 | 0.90 | 0.80 | 0.88 | 0.42 | 0.86 | 0.85 | 0.79 |
| | STDEV | 0.13 | 0.37 | 0.23 | 0.33 | 0.26 | 0.14 | 0.26 | 0.32 | 0.37 |
| Co-occ. semplici (FR) | AVG | 0.89 | 0.86 | 0.86 | 0.63 | 0.88 | 0.85 | 0.82 | 0.64 | 0.59 |
| | STDEV | 0.20 | 0.22 | 0.31 | 0.33 | 0.20 | 0.14 | 0.25 | 0.29 | 0.33 |
| Co-occ. figurate (FR) | AVG | 0.80 | 0.79 | 0.80 | 0.73 | 0.81 | 0.75 | 0.79 | 0.64 | 0.69 |
| | STDEV | 0.35 | 0.36 | 0.37 | 0.30 | 0.33 | 0.00 | 0.34 | 0.33 | 0.28 |
| Co-occ. idiomatiche immediate (FR) | AVG | 0.85 | 0.81 | 0.81 | 0.68 | 0.81 | 0.89 | 0.79 | 0.75 | 0.60 |
| | STDEV | 0.28 | 0.29 | 0.36 | 0.28 | 0.33 | 0.13 | 0.31 | 0.29 | 0.60 |
| **AVERAGE** | AVG | 0.85 | 0.82 | 0.82 | 0.68 | 0.83 | 0.83 | 0.80 | 0.68 | 0.63 |
| | STDEV | 0.28 | 0.29 | 0.35 | 0.30 | 0.29 | 0.09 | 0.30 | 0.30 | 0.41 |

Table 1: Results of the human assessment task for translations, definitions and usage examples. The top half of Table reports results obtained by employing Gemini, and the bottom half table provides the results obtained with Mistral. In each sub-table, we provide average figures and standard deviations obtained by testing on the English translation task, while results on French are at the bottom. We report the results obtained through the three types of considered MWEs: *co-occorrenze semplici*, *co-occorrenze figurate*, and *co-occorrenze idiomatiche immediate* (respectively Literal, Figurative and Immediate Idiomatic MWEs - *supra* pp. 6-7). Detailed results are reported for the translation criteria (1. Correctness, 2. Accuracy and Functional Equivalence, 3. Class Adequacy, 4. Context Usage Suitability, 5. Completeness or Superabundance, 6. Enhancement). The two rightmost columns provide the results of the assessment of definitions and usage examples.

In comparing Gemini and Mistral with respect to English, Gemini achieved slightly higher scores in translation, definition generation, and example creation (averaged over the three types of MWEs, Gemini obtained 0.89 and 0.83 for definitions and examples, respectively; Mistral obtained 0.85 and 0.79). For French, however, the scores are more closely aligned, with a slight advantage for Mistral: averaged ratings amount to 0.66 (definitions) and 0.61 (examples) for Gemini, and to 0.68 and 0.63 for Mistral.

**MISTRAL evaluates GEMINI**

| | | TRANSLATION | | | | | | | DEFINITION | EXAMPLES |
|---|---|---|---|---|---|---|---|---|---|---|
| | METRIC | 1 | 2 | 3 | 4 | 5 | 6 | AVG(1,5) | | |
| Co-occ. semplici (EN) | AVG | 0.98 | 0.80 | 0.80 | 0.83 | 0.90 | *unnec.* | 0.86 | 0.92 | 0.97 |
| | STDEV | 0.16 | 0.26 | 0.26 | 0.25 | 0.24 | *unnec.* | 0.24 | 0.17 | 0.11 |
| Co-occ. figurate (EN) | AVG | 0.97 | 0.83 | 0.86 | 0.88 | 0.87 | 0.92 | 0.88 | 0.98 | 1.00 |
| | STDEV | 0.16 | 0.31 | 0.29 | 0.29 | 0.29 | 0.14 | 0.27 | 0.07 | 0.00 |
| Co-occ. idiomatiche immediate (EN) | AVG | 0.99 | 0.79 | 0.79 | 0.81 | 0.91 | *unnec.* | 0.86 | 0.94 | 0.97 |
| | STDEV | 0.08 | 0.30 | 0.31 | 0.30 | 0.24 | *unnec.* | 0.24 | 0.19 | 0.11 |
| **AVERAGE** | AVG | 0.98 | 0.81 | 0.82 | 0.84 | 0.89 | 0.92 | 0.87 | 0.95 | 0.98 |
| | STDEV | 0.13 | 0.29 | 0.29 | 0.28 | 0.26 | 0.14 | 0.25 | 0.14 | 0.07 |
| Co-occ. semplici (FR) | AVG | 0.98 | 0.84 | 0.84 | 0.88 | 0.90 | *unnec.* | 0.89 | 0.92 | 0.95 |
| | STDEV | 0.16 | 0.23 | 0.22 | 0.20 | 0.20 | *unnec.* | 0.20 | 0.15 | 0.12 |
| Co-occ. figurate (FR) | AVG | 0.98 | 0.82 | 0.83 | 0.85 | 0.90 | 1.00 | 0.87 | 0.97 | 0.98 |
| | STDEV | 0.16 | 0.29 | 0.30 | 0.30 | 0.21 | 0.00 | 0.25 | 0.11 | 0.09 |
| Co-occ. idiomatiche immediate (FR) | AVG | 0.97 | 0.79 | 0.80 | 0.82 | 0.81 | 1.00 | 0.84 | 0.95 | 0.97 |
| | STDEV | 0.12 | 0.30 | 0.29 | 0.29 | 0.31 | 0.00 | 0.26 | 0.12 | 0.11 |
| **AVERAGE** | AVG | 0.98 | 0.82 | 0.82 | 0.85 | 0.87 | 1.00 | 0.87 | 0.95 | 0.97 |
| | STDEV | 0.15 | 0.27 | 0.27 | 0.26 | 0.24 | 0.00 | 0.24 | 0.12 | 0.11 |

**GEMINI evalutes MISTRAL**

| | | TRANSLATION | | | | | | | DEFINITION | EXAMPLES |
|---|---|---|---|---|---|---|---|---|---|---|
| | METRIC | 1 | 2 | 3 | 4 | 5 | 6 | AVG(1,5) | | |
| Co-occ. semplici (EN) | AVG | 0.99 | 0.97 | 0.99 | 0.98 | 1.00 | 0.99 | 0.99 | 0.93 | 0.94 |
| | STDEV | 0.13 | 0.04 | 0.12 | 0.00 | 0.21 | 0.06 | 0.10 | 0.20 | 0.18 |
| Co-occ. figurate (EN) | AVG | 0.98 | 0.79 | 0.86 | 0.81 | 0.98 | 1.00 | 0.88 | 0.83 | 0.88 |
| | STDEV | 0.16 | 0.35 | 0.31 | 0.33 | 0.09 | 0.00 | 0.25 | 0.31 | 0.26 |
| Co-occ. idiomatiche immediate (EN) | AVG | 1.00 | 0.80 | 0.88 | 0.84 | 0.95 | 0.94 | 0.89 | 0.87 | 0.89 |
| | STDEV | 0.00 | 0.31 | 0.29 | 0.31 | 0.18 | 0.22 | 0.22 | 0.25 | 0.25 |
| **AVERAGE** | AVG | 0.99 | 0.85 | 0.91 | 0.88 | 0.98 | 0.98 | 0.92 | 0.88 | 0.90 |
| | STDEV | 0.10 | 0.23 | 0.24 | 0.21 | 0.16 | 0.09 | 0.19 | 0.25 | 0.23 |
| Co-occ. semplici (FR) | AVG | 1.00 | 0.84 | 0.93 | 0.89 | 0.93 | 1.00 | 0.92 | 0.83 | 0.93 |
| | STDEV | 0.00 | 0.25 | 0.16 | 0.23 | 0.20 | 0.00 | 0.17 | 0.28 | 0.15 |
| Co-occ. figurate (FR) | AVG | 0.99 | 0.73 | 0.76 | 0.74 | 0.85 | 0.97 | 0.81 | 0.84 | 0.93 |
| | STDEV | 0.04 | 0.42 | 0.42 | 0.42 | 0.32 | 0.09 | 0.32 | 0.33 | 0.19 |
| Co-occ. idiomatiche immediate (FR) | AVG | 0.99 | 0.84 | 0.90 | 0.87 | 0.96 | 0.97 | 0.91 | 0.89 | 0.94 |
| | STDEV | 0.04 | 0.31 | 0.25 | 0.29 | 0.18 | 0.12 | 0.21 | 0.25 | 0.16 |
| **AVERAGE** | AVG | 0.99 | 0.80 | 0.86 | 0.83 | 0.91 | 0.98 | 0.88 | 0.85 | 0.93 |
| | STDEV | 0.03 | 0.33 | 0.28 | 0.31 | 0.23 | 0.07 | 0.24 | 0.29 | 0.17 |

Table 2: Results of the cross-model assessment task for translations, definitions and usage examples. The top half of Table reports results obtained by employing Mistral to assess Gemini, and the bottom half Table provides the results obtained by assessing Mistral through Gemini. In each sub-table, we provide average figures and standard deviations obtained by testing on the English translation task, while results on French are at the bottom. We report the results obtained through the three types of considered MWEs (see Table 1). Detailed results are reported for the translation criteria (1. Correctness, 2. Accuracy and Functional Equivalence, 3. Class Adequacy, 4. Context Usage Suitability, 5. Completeness or Superabundance, 6. Enhancement), while the two rightmost columns provide results for the definition's assessment and usage examples.

Table 2 shows results of the cross-model assessment task, in which the two models evaluated each other's output. While the scores given by Mistral in evaluating Gemini's output for English are similar to those assigned by lexicographers (across translation, definition, and example tasks), Mistral assigns significantly higher scores for French, especially in the areas of definitions and examples. A similar discrepancy is overall observed in the case of French, where the gap between human judgments about Gemini's output (average scores: 0.81 for translations, 0.66 for definitions, and 0.61 for examples) and Mistral's evaluations (0.87, 0.95, and 0.97, respectively) is even more

pronounced. This comparison suggests that the model struggles to recognize shortcomings in the generation of definitions, and even more in the creation of examples, where the divergence between human and model evaluations is greatest.

| | **Pearson's $r$ (p-value)** |
|---|---|
| English: correlation between human and Mistral, assessing Gemini | 0.824 (0.086) |
| French: correlation between human and Mistral, assessing Gemini | 0.357 (0.555) |
| English: correlation between human and Gemini, assessing Mistral | 0.872 (0.054) |
| French: correlation between human and Gemini, assessing Mistral | 0.511 (0.379) |

Table 3: Correlation coefficients between human ratings and the scores provided by the models on translations. Correlations were computed by considering the five criteria employed to assess the quality of translations: 1. Correctness, 2. Accuracy and Functional Equivalence, 3. Class Adequacy, 4. Context Usage Suitability, 5. Completeness or Superabundance.

A more systematic inspection of the scores characterizing the translations show that the Pearson's correlations between human evaluators and the models are high for English translations, while they are significantly lower for French. Detailed values along with p-values, are provided in Table 3: these show a weaker correlation of both models and thus a reduced agreement with human judgments on French translations, whilst high figures feature a strong agreement for the English translations.

Mistral frequently assigned the *unnec.* value to the Enhancement category (please refer to column 6 in Table 2), including all instances of *co-occorrenze semplici* (or Literal MWEs, in both English and French) and on English *co-occorrenze idiomatiche immediate* (i.e. Immediate Idiomatic MWEs).

### 4.4. Discussion

When evaluating these MWEs, it immediately became clear that their categorisation was not consistent. We could find expressions of different types within one category and, on the other hand, specialised dictionaries labelling the same MWE under different categories. For instance, some expressions taken from letter 'A' are listed in both Tiberii (2012) and De Mauro-Internazionale - thus considered in the first case as *collocazione* and in the second case as *poliremática* (e.g., *acqua e sapone* > EN 'natural beauty'). Similar inconsistencies are present in all the three categories: trying to map these MWEs onto the new overmentioned categories, so as to work according to the scale of metaphoricity functional for our purposes, some readjustments are therefore necessary.

By closely examining the MWEs at hand, we realized that the 120 MWEs are not evenly arranged into our three new categories, but rather 39 *co-occorrenze semplici*, 53 *co-occorrenze figurate*, 10 *co-occorrenze idiomatiche immediate* (*supra* pp. 6-7), and

then 6 expressions with a pragmatic-communicative value; 8 proverbs; 1 cultureme; 3 expressions that are not MWEs. As a consequence, our new categories, *co-occorrenze semplici*, *figurate*, *idiomatiche immediate*, allowed us to rearrange Italian MWEs according to one practical criterion, the degree of metaphoricity, and, consequently, to be able to analyze translations provided by the two models, Gemini and Mistral, within more homogeneous categories. Such homogeneity may facilitate the task of tracking translation issues directly related to LLMs and not to merely linguistic and lexicological ones. In other words, analyzing where LLMs perform best within a set of expressions sharing specific characteristics can help identify key factors that influence translation quality when translating Italian MWEs into French and English.

We observed that the translation quality of the two models is better in cases where:

- there is a greater internal cohesion: e.g. proverbs are generally rendered more correctly than other categories (see *gallina dalle uova d'oro* > EN 'the goose that lays the golden egg', FR 'la poule aux œufs d'or'); the same applies to expressions with a pragmatic-communicative value (*buon appetito!* > EN 'enjoy your meal!', FR 'bon appétit');
- expressions indicating a concrete object or concept (*pianta da appartamento* > EN 'houseplant', FR 'plante d'intérieur'); or in case of specialised technical language (*ritenuta d'acconto* > EN 'withholding tax', FR 'acompte provisionnel').

On the contrary, the translation quality drops:

- when human interpretation is essential, e.g. when quantification is required (*rovescio d'acqua*, EN 'downpour', is certainly not a FR 'déluge' / IT *diluvio*, but it is better translated as 'averse torrentielle'); or in the case of *avere il cuore d'oro* (EN 'to have a heart of gold') translated in French as if it was a complement of matter, 'en or', when the correct expression is 'avoir le cœur d'or';
- in cases of meaning nuances: *accordo di integrazione* is in French rather a 'contrat d'intégration' than an 'accord';
- in the process of rendering expressions linked to the culture of a specific country such as culturemes, and therefore untranslatable (see *Zecchino d'Oro*[15]).

We have also noticed that LLMs tend to grasp the concrete and literal meanings (e.g. *acqua diretta* becomes in French 'eau à la bouche' (Mistral) and 'eau de source' (Gemini)[16]. The further we move away from concrete reality, the more complicated it becomes for models to grasp meanings, unless the expressions are so fixed that they allow for a single possible translation.

[15] The cultureme *Zecchino d'Oro* is an Italian annual competition dedicated to children's music.

[16] In Italian, the French expression 'eau à la bouche' means *acquolina in bocca*; and in English, 'mouth watering'; while, *acqua diretta* is EN 'direct water' and FR 'eau courante'.

## 5. Conclusions

This study investigated how accurately LLMs like Gemini and Mistral generate key components of an online dictionary (translations, definitions, and usage examples) and to what extent their evaluations of linguistic quality align with human judgment. A novel theoretical framework refining the classification of Italian MWEs was introduced, and employed to assess automatically generated translations, definitions, and usage examples. Models' outputs were rated by human lexicographers and in a cross-model setting, whereby models evaluated each other's output.

The experiment revealed language-specific performance differences. For English, the models demonstrated strong accuracy in translating the selected expressions, as well as in generating definitions and usage examples. Their evaluations of each other's outputs also showed high agreement with human judgments. In contrast, the results for French indicated lower translation quality, and the models' assessments of each other's outputs exhibited only modest correlation with human evaluations. This disparity suggests that the reliability of automatic MWE translation may currently be limited to English, or at the very least, still requires critical human supervision, even in resource-rich languages like French. These findings raise concerns about the generalizability of current LLMs across languages when it comes to handling idiomatic, collocational, or otherwise non-compositional expressions.

This line of research may contribute to ongoing discussions in digital lexicography, particularly regarding the automatic creation of dictionary content and the exploitation of language resources. Generative AI may provide helpful tools to lexicography, in particular speeding up the dictionary-making process; however, understanding its capabilities and limitations will be a key factor for developing reliable, high-quality language resources that effectively meet the needs of diverse end users.

The quantitative analysis on English and French translations can be considered as an exploratory study, whose limitations stem from the size of the analyzed data and from the number of annotators. However, this work introduces a sound and replicable protocol to collect MWEs, to analyze them (the number of the models involved can be easily extended), and to compute descriptive statistics to assess the results of the translation processes. Future work will be focused to extend the coverage of the considered MWEs, and to compare further models to analyze their performance in order to identify those most suitable for translating MWEs.

## 6. References

Burger, H. (2010). Phraseologie. Eine Einführung am Beispiel des Deutschen,4., neu bearbeitete Auflage (Grundlagen der Germanistik 36). Berlin: ErichSchmidt.

Giouli, V. & Barbu Mititelu, V. (2024) (eds.). *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives. (Phraseology and Multiword Expressions 6).* Berlin: Language Science Press. Available at: https://langsci-press.org/catalog/book/440

Hjelmslev, L. (1968). *I fondamenti della teoria del linguaggio*. Torino: Einaudi.

Konecny, C. (2010). Kollokationen. Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand italienischer Beispiele. München: Martin Meidenbauer [Forum Sprachwissenschaften; 8].

Marello, C. (1996). *Le parole dell'italiano. Lessico e dizionari*. Bologna: Zanichelli.

Ramisch, C. (2023). *Multiword expressions in computational linguistics*. Computer Science [cs]. Aix Marseille Université (AMU).

Abel, A. (2012). Dictionary writing systems and beyond. In S. Granger & M. Paquot (eds.), *Electronic Lexicography.* Oxford, online edn, Oxford Academic, 24 Jan. 2013. Available at: doi.org/10.1093/acprof:oso/9780199654864.003.0005

Calzolari, N. *et al.* (2002). Towards best practice for multiword expressions in computational lexicons. In M. G. Rodríguez & C. P. S. Araujo (eds.), *Towards Best Practice for Multiword Expressions in Computational Lexicons*. LREC, pp. 1934–1940.

De Mauro, T. (1999). *Introduzione*. In GRADIT 1999-2007, vol. 1º, pp. VII-XLII.

Faloppa, F. (2011). Modi di dire. In S. Raffaele (ed.), *Enciclopedia dell'Italiano (EncIt)*, Roma, Istituto della Enciclopedia italiana. Available at: https://www.treccani.it/enciclopedia/modi-di-dire_(Enciclopedia-dell'Italiano)/

Masini, F. (2019). Multi-Word Expressions and Morphology. In *Oxford Research Encyclopaedia of Linguistics*. Oxford, Oxford University Press.

Voghera, M. (2004). Polirematiche. In M. Grossmann & F. Rainer (eds.), *La formazione delle parole in italiano*. Tübingen: Niemeyer, pp. 56–69.

Garcia, M. *et al.* (2019). Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics. In I. Kosem & T. Zingano Kuhn (eds.). *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 Conference*, 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 747–762. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_42.pdf

Lee, N. *et al.* (2025). Evaluating the Consistency of LLM Evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 10650-10659).

Orenha-Ottaiano, A. (2017). The Compilation of an Online Corpus-Based Bilingual Collocations Dictionary: Motivations, Obstacles and Achievements. In *Proceedings of E-Lex Conference 2017*, Leiden, The Netherlands, pp. 458–473. Available at: https://elex.link/elex2017/wpcontent/uploads/2017/09/paper27.pdf

Orenha-Ottaiano, A. *et al.* (2021). Corpus-based Methodology for an Online Multilingual Collocations Dictionary: First Steps. In Kosem, I., Cukr, M., Jakubíček, M., Kallas, J., Krek, S. & Tiberius, C. (eds.), *Proceedings of*

*Electronic Lexicography in the 21st Century Conference*, 2021-July, pp. 1–28. Available at: https://elex.link/elex2021/wp-content/uploads/eLex_2021-proceedings_compressed.pdf

Sag, I. *et al.* (2002). Multiword Expressions: a Pain in the Neck for NLP. In Gelbukh, A. (ed.), *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Lecture Notes in Computer Science, vol. 2276. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45715-1_1, pp. 1–15.

Gantar, P. *et al.* (2019). Multiword expressions: between lexicography and NLP. In *International Journal of Lexicography*, Vol. 32, n. 2, pp.138–162.

Greco, A. (in preparation), "Una rivisitazione di alcune categorie di 'combinazioni di parole': alcuni criteri lessicografici per la compilazione del Dizionario Nativo Digitale (DND)".

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanit Soc Sci Commun* 11, 426, https://doi.org/10.1057/s41599-024-02889-7

Lew, R. *et al.* (2024). The effectiveness of ChatGPT as a lexical tool for English, compared with a bilingual dictionary and a monolingual learner's dictionary. *Humanities and Social Sciences Communications*, 11(1), pp. 1-10.

Mohammed, T. A. (2025). Evaluating Translation Quality: A Qualitative and Quantitative Assessment of Machine and LLM-Driven Arabic–English Translations. *Information*, 16(6), 440.

Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, 4(1), 100089.

Villavicencio, A. *et al.* (2005) (eds). Special issue on multiword expressions: Having a crack at a hard nut. In *Computer Speech & Language*, Volume 19, Issue 4, pp. 365–377.

**Websites:**

*The Collins Online Dictionary*. Accessed at: https://www.collinsdictionary.com/
*Larousse*. Accessed at: https://www.larousse.fr/
*Le Grand Robert*. Accessed at: https://www.lerobert.com/
*Le Petit Robert*. Accessed at: https://www.lerobert.com/
*Le Robert Dico en ligne*. Accessed at: https://dictionnaire.lerobert.com/
*The Merriam-Webster*. Accessed at: https://www.merriam-webster.com/
The *Oxford Learner's Dictionaries* (learner's dictionaries). Accessed at: http://www.oxfordlearnersdictionaries.com
*Le Trésor de la langue française* (TLFi). Accessed at: http://atilf.atilf.fr/

**Dictionaries:**

De Mauro Internazionale = De Mauro, T., *Dizionario di italiano*

(dizionario.internazionale.it/).
Lapucci, C. (2007), *Dizionario dei proverbi italiani*. Milano: Mondadori.
Russo, D. (2010), *Modi di dire. Lessico italiano delle collocazioni*. Roma: Aracne.
Tiberii, P. (2012), *Dizionario delle collocazioni*. Bologna: Zanichelli.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.
http://creativecommons.org/licenses/by-sa/4.0/

# Annex I - Translations of headwords and MWEs in French[17] and English[18]

### *Collocazioni* as in Tiberii 2012[19]

| Headword (IT) | Headword (EN) | Italian | English |
|---|---|---|---|
| abbonamento | subscription | canone di abbonamento | subscription fee |
| | | fuori abbonamento | ?excluded from subscription |
| abbraccio | embrace | ultimo abbraccio | ?final embrace |
| | | baci e abbracci | ?hugs and kisses |
| abilità | ability | abilità fuori dal comune | ?extraordinary skills/ability |
| | | abilità superiore alla norma | ?above average ability |
| accento | emphasis | accento polemico | ?polemical tone |
| | | accento di rimprovero | ?tone of reproach/reproachful tone |
| acconto | deposit | ritenuta d'acconto | withholding tax |
| accordo | agreement | mutuo accordo | mutual agreement |
| | | accordo aziendale | collective bargaining agreement |
| | | agire in accordo | ?act in concert |
| | | andare d'accordo | get along |
| | | andare d'amore e d'accordo | ?to get along very well |
| | | essere tutti d'accordo | ?to be all in agreement |
| | | accordo di integrazione | ?integration agreement |
| | | accordo di riservatezza | non-disclosure agreement |
| acqua | water | acqua da bere | easy peasy |
| | | rovescio d'acqua | downpour |
| | | sorso d'acqua | ?a sip/drink of water |

[17] For French, contact authors.

[18] A question mark preceding our English translations indicates that we found the MWEs in several reliable sources on the web, but not in a dictionary.

[19] We report here the original lists of expressions, and their official translations in English and French, taken from specialised dictionaries (see 3.1).

| | | | |
|---|---|---|---|
| | | via d'acqua | waterway |
| affare | deal | grosso affare | big deal |
| | | vero affare | real bargain |
| | | affare garantito | sure thing |
| | | affare illecito | shady deal |
| | | affare illegale | illegal deal |
| | | affare immobiliare | ?real estate deal |
| | | affare rischioso | ?risky business |
| | | affare sospetto | ?fishy/shady deal |
| | | affare vantaggioso | a great/good deal' = a lot of; AmE: sweetheart deal |
| | | avere un affare per le mani | ?have a big deal going down |
| | | essere coinvolto in un affare | ?be involved in a deal |
| | | fare affari d'oro | strike gold |
| | | mandare a monte un affare | scupper a deal |
| | | prendere parte a un affare | ?take part in a deal |
| | | un affare va a buon fine | a deal goes through |
| | | affari a rischio | risky business |
| | | affare ambiguo | ?dodgy deal |
| barba | beard | lozione da barba | aftershave lotion |
| bellezza | beauty | bellezza acqua e sapone | ?natural beauty |

***Espressioni polirematiche* as in De Mauro-Internazionale**

| **Headword (IT)** | **Headword (EN)** | **Italian** | **English** |
|---|---|---|---|
| abbonamento | subscription | polizza di abbonamento | ?insurance for safe deposit box |
| aceto | vinegar | aceto inglese | ?smelling salts |
| | | aceto balsamico | balsamic vinegar |
| | | aceto di mele | ?apple vinegar |
| | | aceto di vino | wine vinegar |
| | | sott'aceto | pickled |
| acqua | water | a pane e acqua | ?on bread and water |
| | | acqua alta | ?high water |
| | | acqua benedetta | holy water |
| | | acqua cheta | still waters run deep |
| | | acque bianche | ?white water |
| | | acqua diretta | ?direct water |
| | | acque territoriali | territorial waters |
| | | acque luride | blackwater |
| | | acque reflue | wastewater |
| | | acque freatiche | groundwater |
| | | buco nell'acqua | washout |
| | | colore ad acqua | watercolor |
| | | imbarcare acqua | ?take on water |
| | | acqua passata | water under the bridge |
| | | fare acqua | leak |
| | | acqua fresca | chit(-)chat |
| | | acqua in bocca | mum's the word |
| aereo | aeroplane | spazio aereo | airspace |
| | | legante aereo | ?air binder |
| stampa | press | addetto stampa | press officer |
| affari | deals | giro di affari | turnover |
| alimentari | food | generi alimentari | foodstuff |

| | | | |
|---|---|---|---|
| alimentazione | feeding | alimentazione forzata | force-feeding |
| anello | ring | anello di fidanzamento | engagement ring |
| | | anello nuziale | wedding ring |
| angolo | corner | calcio d'angolo | corner kick |
| animale | animal | animale ragionevole | ?rational animal |
| anno | year | capo d'anno | new year's day |
| | | fiore degli anni | ?prime of life |
| | | perdere l'anno | ?fail the year |
| anticipo | early | in anticipo | ?ahead of time |
| appartamento | apartment | pianta da appartamento | houseplant |
| appetito | hunger | buon appetito | enjoy your meal |
| aprile | April | pesce d'aprile | April fool's joke |

***Modi di dire* as in Russo (2010)**

| **Headword (IT)** | **Headword (EN)** | **Italian** | **English** |
|---|---|---|---|
| oro | gold | anello d'oro | golden ring |
| | | non è sempre oro quel che luccica | all that glitters is not gold |
| | | celebrare un oro | ?celebrate a gold |
| | | pagare qualcosa oro | pay an arm and a leg |
| | | Zecchino d'oro | ?Zecchino d'Oro |
| | | avere il cuore d'oro | Heart of gold |
| | | per tutto l'oro del mondo | ?for all the money in the world |
| | | albo d'oro | hall of fame |
| | | sogni d'oro | sweet dreams |
| | | gallina dalle uova d'oro | golden goose |
| | | adorare il vitello d'oro | ?to worship the golden calf |
| anima | soul | fare l'anima bella | a wolf in sheep's clothing |
| | | essere l'anima della compagnia | the life and soul of the party/? the life of the party |
| | | buttarsi anima e corpo | throw oneself body and soul |
| | | non esserci anima viva | ?there wasn't a soul in sight |
| | | mettersi l'anima in pace | ?found peace in their soul |
| | | rendere l'anima a Dio | ?give up the ghost |
| | | tenere l'anima con i denti | ?to cling on for dear life |
| | | vendere l'anima al diavolo | sell your soul (to the devil) |
| | | all'anima | ?wow |
| | | buon'anima | ?The late [Name] |
| | | in corpo e in anima | ?heart and soul |
| | | con la morte nell'anima | with a heavy heart |
| anello | ring | anello del vescovo | ?bishop ring |
| | | anello di una pista automobilistica | racetrack |
| | | anello di una catena | link in the chain |
| bello | beautiful | sarebbe bello che piovesse | ?it would be nice if it rained |
| | | il bello della campagna è il silenzio | ?the beauty of the countryside |

| | | | |
|---|---|---|---|
| | | trovare una bella soluzione | ?find a good solution |
| | | non è bello dire certe cose | ?it's not mice to say such things |
| | | ricatto bello e buono | ?plain and simple blackmail |
| | | troppo bello per essere vero! | too good to be true |
| | | fare il bello e il cattivo tempo | ?call the shots |
| | | l'amore non è bello se non è litigarello! | love isn't beautiful if it isn't a bit quarrelsome! |
| | | raccontarne delle belle | ?tell tall tales |
| | | farne di belle e di brutte | ?to get up to all sorts of mischief |
| bocca | mouth | avere il miele sulla bocca e il veleno nel cuore | ?frenemy |
| dare | to give | dare le pecore in guardia al lupo | ?give the sheep to the wolf to guard |
| | | dare a Cesare quel che è di Cesare | ?render unto Caesar what is Caesar's |
| | | dare un colpo al cerchio e uno alla botte | straddle the fence |

# Annex II - Employed Prompts

### Prompt for the translation generation

The following prompt was employed to obtain the translation, definition and usage example for each considered ***input expression***.

*You are a professional translator. I will provide an Italian expression, and you must return its most accurate English translation along with a clear usage example and a brief explanation of its meaning.*
*Carefully consider the best translation and ensure your response follows this format: Translation: <your translation>. Example: <a single usage example in English>.*
*Explanation: <a brief explanation of the expression's meaning>.*
*Provide only one translation. Expression: '{**input expression**}'.*

### Prompt for the assessment of the translation

The following prompt was employed to ask a model to assess the quality of a ***translation*** for an input ***expression***, based on the specified ***correct_translation***.

*Evaluate the quality of the following Italian-to-{target_language} translation on a scale from 1 to 5, where 1 is completely incorrect and 5 is perfect:*

> *Italian expression: {**expression**}*
> *{target_language} translation: {**translation**}*
> *correct translation: {**correct_translation**}*

*Consider: Correctness, Accuracy and Functional Equivalence, Class Adequacy, Context Usage Suitability, Completeness or Superabundance, Enhancement, with the given definitions:*

*Correctness: the rendering in the target language is clear and correct in spelling, morphology and syntax.*

*Accuracy and Functional Equivalence: the message from the source language to the target language is the same: meaning, communicative intention and connotation are maintained.*

*Class Adequacy: there is an adaptation of category and in gradation of meaning.*

*Context Usage Suitability: the usage is the same in both languages.*

*Completeness or Superabundance: information is either complete or missing or there is additional information.*

*Enhancement: there are signs of refinement of meaning in the target language rendering, such as punctuation marks, capitalization where necessary, etc.*

*Provide only a json with a number from 1 to 5 for each element of evaluation. No other text is needed.*

### Prompt for the assessment of the definition and of the usage example

The following prompt was employed to ask a model to assess the quality of a ***definition*** and of an ***example_of_usage*** for an input ***expression***, based on the specified ***translation*, *correct_translation***.

*Evaluate the quality of the definition and the example of usage of the following Italian-to-{target_language} translation on a scale from 1 to 5, where 1 is completely incorrect and 5 is perfect:*

> *Italian expression: {**expression**}*
> *{target_language} translation: {**translation**}*
> *correct translation: {**correct_translation**}*
> *definition: {**definition**}*
> *example of usage: {**example_of_usage**}*

*Provide only a json with a number from 1 to 5 for definition and usage_example. No other text is needed.*

## Annex III - Table of Categories

| Levels of Metaphoricity | Definitions in Italian | Definitions in English | Example in Italian | Equivalent in English[20] |
|---|---|---|---|---|
| 1st level | Co-occorrenze libere | [not analysed here] | [not analysed here] | [not analysed here] |
| 2nd level | Co-occorrenze semplici | Literal MWEs | *agenzia di viaggi* | 'travel agency' |
| 3rd level | Co-occorrenze figurate | Figurative MWEs | *tenere aggiornati* | 'to keep someone informed about' |
| 4th level | Co-occorrenze pragmatico-comunicative | [not analysed here] | [not analysed here] | [not analysed here] |
| 5th level | Co-occorrenze idiomatiche immediate | Immediate idiomatic MWEs | *avere il cuore d'oro* | 'to have a heart of gold' |
| 6th level | Co-occorrenze idiomatiche mediate | [not analysed here] | [not analysed here] | [not analysed here] |
| 7th level | Co-occorrenze proverbiali o di metafora con altre realtà | [not analysed here] | [not analysed here] | [not analysed here] |

[20] Please, note that English equivalents do not reflect the semantic distribution of meanings in Italian. So, we urge you to not apply Italian categories to English examples but following our examples and descriptions of Italian MWEs. As mentioned above, the semantic content (or meaning) of MWEs is only sometimes distributed equally between the two languages considered here, Italian and English.