International Journal Of Lexicography, 2026, 39, 1–26
https://doi.org/10.1093/ijl/ecag017
Published: 28 July 2026
Article

OXFORD

Extraction and abstraction of lexicographic data through AI: The case of Body Part Nouns

Laura Giacomini, and Valentina Piunno

University of Innsbruck, Austria (laura.giacomini@uibk.ac.at)
University of Bergamo, Italy (valentina.piunno@unibg.it)

Abstract

This paper aims at exploring the use of Artificial Intelligence for lexicographic purposes. Focusing on a specific domain that is underrepresented in lexicographic works - namely, the domain of Body Part Nouns - the contribution employs AI to i) extract phraseological expressions and authentic examples in context, and to ii) organize the data according to specific properties, such as semantic domains and cognitive processes involved. The paper describes both the prompt-construction phase and the qualitative evaluation of the results obtained. Extraction tasks are compared with analysis tasks in order to identify the areas in which the LLM model achieves better performance and those where it still has limitations. While the primary focus of this contribution is on Italian data, the study also offers some contrastive insights into English.

Keywords phraseological expressions; semantic domains; AI for lexicography; Body Part Nouns

1. Body part nouns: between lexicology and cognition[1]

Since Wittgenstein's work ([1953]1958), scholars have described from different perspectives how the meaning of words reflects and is deeply shaped by their (social) contexts of use[2]. Within this field, particular attention has been devoted to Body Part Nouns (hereinafter BPNs), which have been examined through diverse approaches and across multiple disciplines, such as linguistics, philosophy of language, anthropological and cognitive sciences. As far as the former is concerned, several areas of lexicology have focused on the role of body part vocabulary. Among them, at least lexical semantics (Goddard, 2002; Wierzbicka, 2007), cognitive semantics (Lakoff & Johnson, 1980; Johnson, 1987; Talmy, 2000; Kövecses,

2000; Ruthrof, 2000) and phraseological studies (e.g. among others, Manerko, 2014; Ganfi et al., 2023) can be mentioned. Within these areas, the role of body parts has been investigated in relation to several features, such as their salience with respect to the human experience, their contribution to the conceptualization of abstract notions and their universality traits.

In the research concerning lexical semantics driven by Wierzbicka (1972, 2007, 2014), the body is one of the limited set of *semantic primes*, namely basic meanings that are not further definable and can be identified in all languages (Wierzbicka, 2007). In this perspective, semantic primes constitute a minimal, accessible and non-ethnocentric "principled 'vocabulary' for semantic-conceptual representation" (Wierzbicka, 2007: 19), which constitute the fundamental building block of more complex semantic units (such as the parts of the body, e.g. *arm, head*, *leg*, etc.). As Wierzbicka (2007: 15) puts it, "the domain of the human body is an ideal focus for semantic typology and, […], cognitive anthropology, because the body is, almost certainly, a *conceptual* […] *universal*" (our italics). Furthermore, despite language variation, speakers of different languages "appear to think about the human body in terms of certain parts" (Wierzbicka, 2007: 15). Scholars have debated on the notion of universality both in terms of the linguistic and conceptual segmentation of the body parts (Brown, 1976, Majid et al., 2006) and in relation to the use of the BPNs for the conceptualization of the world's different notions (as, for instance, topological relationships and spatial values, cf. at least Heine, 1997; Heine & Kuteva, 2002). The former perspective has demonstrated that "linguistic categorisation of the body is subject to both universal and language-specific principles" (Enfield et al., 2006: 145).

The latter topic represents a central aspect of cognitive semantics and of typological and grammaticalization studies (cf. at least Lehmann, [1982]1995; Heine & Kuteva, 2002; Hopper & Traugott, 2003; Koptjevskaja-Tamm et al., 2007; Evans, 2010). On the one hand, the research by Lakoff and Johnson (1980, 1999) focuses on the use of BPNs for the conceptualization and expression of the bodily experience (i.e. the notion of *embodiment*). In this perspective, BPNs are a key feature of the human conceptual system (Kövecses, 2000; Ruthrof, 2000; Evans, 2010), since the embodied experience allows to convey more complex concepts (Evans & Green, 2006: 46). BPNs serve this function properly. On the other hand, it has been demonstrated that, among the world's abstract concepts, the use of BPNs is particularly widespread in the languages of the word to convey grammatical values: "this lexical field is so central to human life that it has its own grammar, and its elements are often grammaticalized to function in the grammar themselves" (Lehmann, 2022: 14).

Recent studies have clearly shown the interconnection between the emerged grammaticalized functions of BPNs and the degree of lexicalization of the multiword units that contain them (Ganfi et al., 2023). This topic is of particular interest to lexical studies devoted to phraseology. The present contribution is situated within this field, drawing on theoretical work in lexicology to shed light on the presentation of the different abstract domains that can be connected to BPNs in lexicographic works (Section 2). To this purpose, we aim to test the potential of Artificial Intelligence for lexicographic purposes (Section 3).

1.1. Cognitive processes and abstract domains

By *body part* we mean any distinct, functional part of a living organism's body, especially in animals and humans, among which not only external parts like the nose or a hand, but also internal parts, even organs, like the heart or the brain. BPNs are polysemous lexemes, in the sense that they show "multiple distinct yet related meanings" (Evans & Green, 2006: 36): despite a core meaning, they convey other senses which "vary from occurrence to occurrence as a

function of the interaction with the other words it combines with" (Ježek, 2023: 226). Take for instance the word *head* in English. It primarily denotes the uppermost part of the (human) body; nevertheless, the lexeme conveys a variety of other meanings, such as the following:

(1) *head of a company* → the person in charge
(2) *head of the table* → the most important seat
(3) *head of lettuce* → a single whole unit of the vegetable

For instance, the extended meaning of (1) and (2) could be related among them (in *head of a company* and *head of the table*, the lexeme *head* represents the apex of a hierarchical scale. Lexicological and typological studies have devoted considerable attention to identifying i) the factors that contributed to the development of semantic extensions[3] and ii) the commonalities between the different values. Two correlated mechanisms are typically involved in the semantic extension of lexemes, and thus in the emergence of lexical polysemy: metaphor and metonymy (cf. Lakoff & Johnson, 1980; Lakoff, 1987). They represent a core feature of human language (Evans & Green, 2006) and "fundamental linguistic mechanisms which regulate the variation in the meaning of units" (Robert, 2008, p. 61). Furthermore, these cognitive processes are often involved in the conceptualisation of human experiences by means of BPNs.

Metaphor occurs when an abstract notion is conceptualized in terms of a more concrete domain (e.g. the hierarchical values of *head* that were mentioned in examples (1)-(2)). Metonymy, on the other hand, relies on a relation of logical contiguity between two distinct domains (e.g. the *head* in example (3) designates a single unit of lettuce, by logical contiguity with the head of the single individual). As some scholars have highlighted, a regular semantic pattern which is common to the whole set of domains can be identified (Robert, 2008; the *semantic super-structure* in the terms of Michaelis, 1993, 1996). The pattern regulates the formation of new senses or forms of the lexeme. How are metaphor and metonymy concretely involved in the development of polysemy? While metaphor is typically represented as a radial shift, where the distinct semantic extensions originate from a core meaning (Lipka, 1992, p. 127), metonymy is a chaining shift, since the semantic development involves a chain of senses derivation (where each sense, in turns, derives from a previous one), cf. Figure 1:

On the one hand, it has been demonstrated that metaphorical and metonymic shifts can be strictly interrelated in the semantic extension of lexemes (Robert, 2008; Balbachan, 2006; Lipka, 1992). On the other hand, several derivational relationships have been identified for

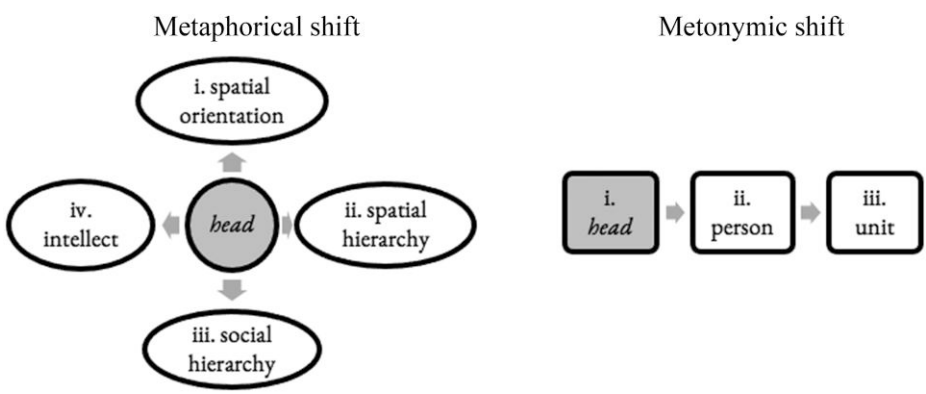

Figure 1 Radial and chained shifts (cf. Piunno, 2023, p. 149).

different BPNs (cf. at least Ganfi et al., 2023). The frequent use of such schemas contributes to the extension of the lexicon by the exploitation of already available resources (Piunno 2023: 173). This is particularly true for phraseological units, which, as demonstrated in Ganfi et al. (2023), often lexicalize these values in a set of different syntagmatic patterns.

The challenge we set out in this paper is to integrate cognitive processes of this kind, which are virtually universal and lie at the very foundation of human cognition, to the use of Artificial Intelligence. Since AI imitates human communication, it should simulate human language in all its nuances, including cognitive processes such as these, which are crucial, particularly when they develop lexical polysemy (with special reference to lexemes denoting body parts), as they establish connections between abstract concepts and cognitive mechanisms. In particular, we intend to use chatbots to identify: (i) examples of phraseological sequences containing BPNs conveying abstract meanings, (ii) the different abstract domains that can be associated with the same body part, and (iii) the cognitive processes involved in each semantic extension. We aim at understanding to what extent AI can facilitate the identification of these phenomena, enabling a more accurate processing and presentation of lexicographic data.

2. Representation and presentation of metaphorical and metonymical body part nouns in lexicography

An examination of major online general dictionaries of Italian, English and German reveals, on the one hand, a lack of—or only partial—presentation of BPNs as a homogeneous semantic field. Tools such as cross-references, topic-oriented disambiguation markers, or nomenclature tables that link BPNs to each other are mostly not available. Some of these features can be found in online learner's dictionaries intended for L2 or foreign language users, as well as in school dictionaries, which is appropriate for their pedagogical aims. LDOCE online (2008), for example, includes BPNs (as well as words more broadly related to the concept) within the wide-ranging onomasiological area *Topic: Human*, thereby offering only a partial solution to this issue. A more coherent solution, similar to those found in picture dictionaries, is offered by OALD online (2020), in which each lexical entry for a BPN is accompanied by the same colour illustration of a man, labelled with the corresponding (external!) body part nomenclature.

On the other hand, the treatment of BPN-related idioms tends to be limited to lists mostly arranged in alphabetical order. CLD online (1999) and *Il Nuovo De Mauro*, for instance, cluster idioms in a distinct section with a distinct structure. While the list in the former is just made up of links to separate entries for each expression, the latter also provides a preview of the definition. A similar solution is adopted by the *Collins Online Dictionary* (2007), which, however, presents a numbered list of idioms that continues the sequence of the main senses of the word, without allocating a dedicated section. The idioms themselves are linked to independent entries. OALD (2020) provides a list of idioms in a side menu; each of them is linked to the entry of the word under which it is recorded. Idioms are therefore not treated as independent entries. For example, in the list of idioms related to the BPN *head*, one finds *have eyes in the back of your head*, which links to the entry for the word *back*. There, the expression is provided with a definition and usage examples within a dedicated section titled *Idioms*.

Not infrequently, idioms are not explicitly marked as such, lack specific usage labels, are presented alongside collocations of the headword, or are even embedded in general usage examples. To the best of our knowledge, no general language or learner's dictionary in Europe provides systematic information concerning, on the one hand, the metaphorical or metonymical nature of a given idiom and, on the other hand, semantic relations between idioms sharing the

same abstract domain. Explicit semantic relations between idiomatic expressions as well as clustering of idiomatic expressions based on such relations are also missing in contemporary lexicography. In light of the great phraseological productivity of BPNs, such presentation formats hinder data accessibility and reduce the overall user-friendliness of the dictionary.

2.1. General requirements: the example of the Phrase-based Active Dictionary

Depending on the type of phraseological unit and the type of dictionary, lexicographic treatment may require different strategies in terms of placement, ordering, and labelling. While collocations are often idiosyncratic and not predictable in text production, they are nonetheless compositional and therefore relatively transparent for receptive purposes (cf. Hausmann, 1984). As a result, they do not usually require a definition, while providing usage examples is beneficial. Idiomatic and semi-idiomatic expressions, by contrast, require a different type of treatment—one that includes at minimum a definition so as to make their non-compositional meaning explicit.

To highlight further microstructural aspects central to the lexicographic presentation of idioms, as well as issues related to their pre-lexicographic acquisition, we will refer to the requirements of a monolingual learner's dictionary model addressing advanced learners and translators: the *Phrase-based Active Dictionary* (PAD). The dictionary model is currently under development and grounded in phraseological and cognitive principles. It is being implemented in German, Italian and English within the framework of the PhraseBase project (cf. Giacomini, 2025). The microstructure of the PAD, described in detail in DiMuccio-Failla & Giacomini (2022), consists of multiple levels of description that allow for the progressive disambiguation of a word's contextual meaning: the level of the syntactic construction, the level of the sense field, and the level of the lexical unit, which forms the core of the lexicographic entry (cf. Figure 2).

The lexical unit consists of a typical and unique context in which the word is embedded and is accompanied by a definition and usage examples. The lexical unit represents a phraseological expression (or a group of related phraseological expressions), and as such, it also serves as an appropriate level of description for fixed expressions. The PAD aims at integrating metaphorical and metonymic idioms within its overall microstructure.

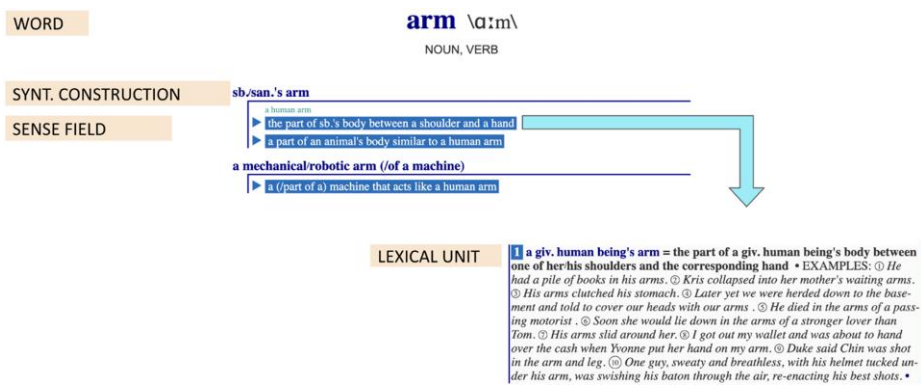

Figure 2 Microstructure of the PAD.

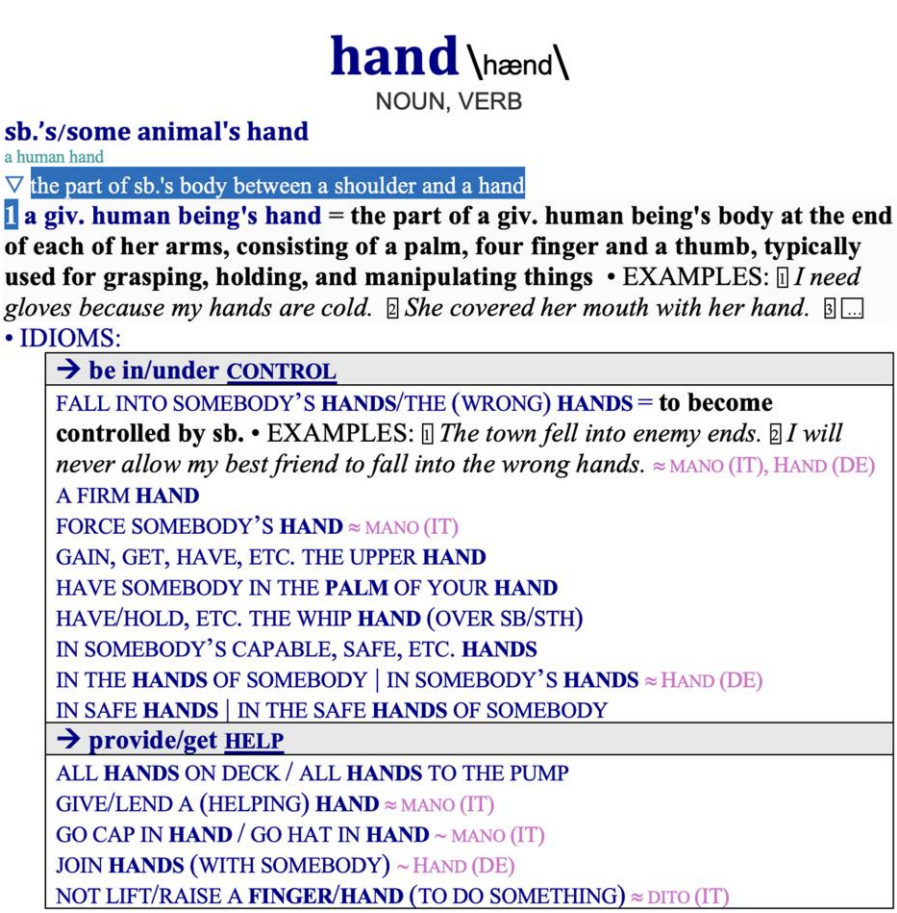

Figure 3 Option for idiom presentation in the English PAD: idioms are clustered around an abstract domain (e.g. CONTROL), followed by a definition, usage examples, and, if relevant, by a cross-reference to similar entries in the German or the Italian PAD.

The PAD provides both syntactic and semantic access to lexical units: syntactic access is granted through syntactic constructions (e.g. *an arm of sth*), which group lexical units according to their structural patterns; semantic access is provided through sense fields (e.g., *an elongated part of a geographical area*), which group lexical units according to a shared basic meaning. The same principle can be applied to idioms, which function as fully-fledged lexical units and, as such, can be matched to a specific syntactic construction and a particular sense field (cf. Figure 3).

The following section introduces and discusses key aspects involved in the treatment of idioms in the PAD, observed during work at BPNs:

- **External syntactic and semantic access to idioms and alignment with the characteristics of a bifunctional (active and passive) dictionary:**

  Although the PAD is primarily designed as an active dictionary, its use for receptive purposes—especially in the case of compositional expressions—is far from negligible. A syntactic grouping enables users to trace a definition by recognizing the construction. Conversely, identifying a cluster of expressions with related meanings presupposes prior knowledge of the expression's underlying meaning.

- **Idioms as items within the BPN entry or as independent entries:**
  Current experiments focus on treating idioms as elements within the entry of the corresponding noun, which—as previously discussed—makes it possible to handle them in accordance with the microstructural principles of the PAD model. Independent entries for idioms would be a possible solution from the perspective of user access, although they would be more useful for text reception and cognitive usage situations than for text production.
- **Use of abstract semantic domains:**
  Semantic grouping can rely not only on general sense fields but also on abstract domains, for instance
  - CONTROL/POWER for *the long arm of sth* (typically: *of the law*) and Italian *il lungo braccio di qs* (typically: *della legge*),
  - DESIRE for *sb would give their right arm to do sth* and Italian *qn darebbe un braccio per poter fare qs*,
  - DISTANCE/AVOIDANCE for *keep/hold sb at arm's length*.

  These serve as more specific disambiguation tools for BPNs and can also facilitate the creation of cross-references among different BPNs, thereby enhancing onomasiological access to the data[4]. For instance,
  - CONTROL is shared by expressions such as English *gain, get, have, etc. the upper hand*, *have somebody in the palm of your hand*, and *have/hold, etc. the whip hand (over somebody/something)*;
  - HELP/SUPPORT is shared by expressions such as *all hands on deck* (also *all hands to the pump*) and *give/lend a helping hand*.
- **Figurative vs. literal meaning:**
  Many expressions based on BPNs have both a figurative and a literal meaning (e.g. English *to twist sb's arm* and Italian *incrociare le braccia* ('to cross one's arms' and 'to go on strike')). This can sometimes lead to a dual, separate treatment of the same expression—e.g., as an idiom and also as a collocation. In such cases, it is essential that users can access both pieces of information, and that an explicit connection is made between them.
- **Classification of metaphor and metonymy:**
  The distinction between metaphor and metonymy is functionally relevant for the representation of data within the PAD database, as it allows for the reconstruction of the cognitive network underlying the meanings of individual words and word groups—one of the central goals of the PAD model (cf. DiMuccio-Failla, 2025). This kind of information can help users understand how metaphorical and metonymical lexical units (i.e., distinct senses) have originated from a core, concrete lexical unit—for example, *arm of a cross* and *arm of a company* from *a human being's arm*—through recurring cognitive mechanisms. Explicit lexicographic labelling of metaphor or metonymy can be included in the dictionary entry.
- **Metaphorical and/or metonymic nature of an idiom:**
  Metaphor and/or metonymy may affect either the entire expression or only the BPN within it. This phenomenon can lead to a further fragmentation of data and may require an additional level of specification in the dictionary structure.

2.2. Challenges of data acquisition

Within the lexicographic process, the acquisition of data relating to idiomatic expressions represents a particularly complex challenge. First of all, identifying the expressions themselves and sorting them is a time-consuming task: in the case of BPNs, their high number typically results in only partial coverage in general and learner's dictionaries, and even phraseological

dictionaries do not always offer satisfactory solutions. In general, lexicographic coverage is quantitatively and qualitatively inadequate. Furthermore, in order to meet the presentation requirements outlined in the previous section, this type of language element must be accompanied by a wide range of information: definitions, usage examples, pragmatic usage labels, indications of the metaphorical or metonymic nature, and references to abstract semantic domains. Furthermore, mediostructural components are needed to connect idiomatic expressions across different BPNs.

To this end, corpus linguistic methods are only partially helpful. Semantic indications as well as groupings based on abstract domains remain one of the greatest challenges—so far addressed primarily through non-automated methods, e.g. introspection or deduction based on available lexicographic data. Lexicographic resources that provide semantic access to phraseological data constitute an exception. Some non-cumulative combinatorial dictionaries offer both syntactic and semantic groupings, but without fully systematizing the underlying methodology (e.g., the use of conceptual tags in some of the entries of REDES (Bosque Muñoz, 2004)).

The online Italian phraseological dictionary GEPHRI (Schafroth et al., 2018), by contrast, allows access to phraseological units by semantic fields. Semantic fields have two levels of granularity: a broad field (e.g., emotional states) and a more fine-grained subfield (e.g. confusion, joy, bad mood, impatience). The level of granularity found in the subfields is comparable to the one of the abstract semantic domains described in this study. This constitutes a notable exception in the present state of lexicographic practice.

3. Discussion

As mentioned in the previous section, traditional corpus linguistic methods are not sufficiently efficient in retrieving the type of information on idioms that is necessary for lexicographic purposes. In particular, a conceptual categorization of idiomatic expressions remains largely inaccessible to traditional methods, due in part to the lack of reliable and comprehensive conceptual (onomasiological) resources for general language. In this regard, even resources such as wordnets—which are available for all the languages examined in this study—are only partially helpful, given the issues documented in the relevant literature (see, among others, Hanks & Ježek 2008). As a result, much of the required work has to be carried out manually, which significantly burdens the lexicographic process in terms of time.

The present study aims to test the potential of generative AI as a tool for collecting and classifying metaphorical and metonymic idiomatic expressions, along with the relevant lexicographic data, in an attempt to automate this task as far as possible while consistently prioritizing the quality of the final result.

Recently, the use of generative artificial intelligence has attracted attention in lexicographic research, precisely because the inherently complex nature of the data acquisition and compilation process could potentially benefit from the new technologies.

De Schryver (2023) explores the evolving role of generative AI in lexicography by reviewing ten early studies that assess its capacity to automate key dictionary-making tasks and calls for a proactive, critical engagement with AI, urging lexicographers to shape its use rather than be displaced by it. Along similar lines, Lew (2023) and Rundell (2023) show that despite the strengths of AI, e.g. generating high-quality dictionary definitions and handling transparent terms, other crucial tasks, e.g. retrieving example sentences and resolving polysemy, are much less satisfactory, concluding that human lexicography remains essential. Both Lew (2024) and Ptasznik et al. (2024) describe similar results obtained by focusing on dictionary functions (reception, production, translation). Some case studies deal with the dictionary as a specific text

type and test the automated generation of entries (Arias-Arias et al., 2024; de Schryver, 2023). The role of (computer-assisted) human post-editing in lexicography, similar to the one experienced in translation studies when dealing with machine translation, is another significant topic in some of these contributions (e.g. Rundell, 2023 and Jakubíček & Rundell, 2023). Post-editing is intended as the evaluation and processing of a first lexicographic draft. Finally, McKean & Fitzgerald (2024) also sheds light on the return of investment of LLMs in lexicography, critically examining aspects such as environmental impact, replication of bias, and hallucinations.

Transparency, traceability, precision and consistency are just some of the relevant criteria required for professional dictionary-making applied by these studies when testing generative AI performance. Together, they underscore the promise of AI in lexicography, while emphasizing the importance of integrating AI with conventional tools and the continued need for expert oversight and ethical use.

While these studies primarily focus on other types of lexical data or full entries, there seems to be a lack of research explicitly dedicated to the processing of phraseological data for lexicography, while initial observations on idiomaticity detection have been made in the fields of Computer Science and Data Science, showing that general-purpose LLMs can be effective at identifying idiomatic language, but task-specific encoder-only models still perform best (cf. Phelps et al., 2024).

At this stage, we will not consider terminology-oriented research. Although it has recently focused even more intensively on this topic, it is only partially aligned with the objectives of our study.

Our goal is to test the performance of AI—specifically ChatGPT—in extracting and sorting idiomatic expressions related to BPNs. In doing so, we place particular emphasis on the role of prompting in retrieving the desired data. Prompt design is a fundamental component of our experiments, which involve a process of progressive refinement of the output in order to identify not only the data most closely aligned with our purposes, but also the most suitable prompt format for replication in similar data acquisition contexts.

Since prompt design significantly affects output quality, White et al. (2023) introduce a catalogue of prompt patterns designed to improve the practice of prompt engineering. In our case study on BPNs, we take into consideration some of the proposed pattern categories as being relevant within the lexicographic process, namely

(a) input semantics, specifically the *creation of metalanguage* pattern (key concepts are explained within the query itself, not least by means of examples),
(b) output customization, specifically the *output automater* pattern and the *template* pattern (to produce structured, and repeatable outputs), and
(c) error identification, specifically the *reflection pattern* (to review and improve the previous answer).

In our experiments, these patterns play a significant role in aiming at prompt effectiveness. We decided not to use prompts involving "impersonation" (cf. persona pattern in White et al., 2023), even though they tend to yield higher performances (Phelps et al., 2024), so as not to lose sight of our main focus: in fact, our aim was not to test the AI on building lexicographic entries, but rather on representing a specific set of expressions and organizing them according to their meanings.

While the possibility of applying an existing catalogue like this is beneficial when crafting prompts for a new area of research, we are aware that, even by initially setting up a prompting framework with specific patterns, it is not possible yet to move beyond trial-and-error

prompting. Nonetheless, our objective remains to approximate, as closely as possible, a reusable model aimed at processing idioms for lexicographic purposes.

In what follows, we describe the components and steps involved in the creation of the prompt (Section 3.1), as well as the structure of the final prompt selected for the analysis and the results obtained (Section 3.2). We also assess a prompt auto-generated by ChatGPT (Section 3.3) and compare the output in Italian with the one in English, the language on which the model was originally trained (Section 3.4).

3.1. Structuring prompts: tasks and expected outputs

To assess the potential use of AI in the field of lexicography, we carried out a series of tests on the functionalities of GPT-5[5]. In what follows, we describe the stages of the process of prompt construction, the results, as well as the system's limitations and potentials.

The effectiveness of the application of AI and of generative models to lexicography largely relies on the quality of prompts. Thus, the accurate engineering of prompts is a key factor in achieving targeted outputs (Short & Short, 2023; Eager & Brunton, 2023; Giray, 2023; Federiakin et al., 2024; He et al., 2024). For the purpose of this analysis, we conducted two types of queries: in the first, the prompts required as its output a set of examples of word combinations including different body parts and grouped by abstract domains; in the second query, the requested output consisted of examples containing a specific body-part term and conveying an abstract meaning. Six different prompts were generated using both a *zero* and a *few-shot model*: in the first set of prompts no prior examples were given, but some fine-tuning instructions were provided; in the second set of prompts, the LLM was instructed by means of a few examples which aimed at illustrating how the task had to be performed[6]. The latter proved to be more suitable for our research, particularly in view of the more complex output we required. All prompts address the precise goal and context (e.g. the creation of a lexicographic entry and the representation of its abstract domains as a network of meanings), the content that is required (e.g. the metaphorical/metonymic uses of a BPN), as well as the output required (e.g. a set of examples). Prompts are aligned with the guidelines outlined by Eager and Brunton (2023: 4-5) and Giray (2023: 2630), and include the following components: i) the action to be performed, ii) the outcome of the action, iii) the context or parameter of the task, iv) focus or conditions for the generated output, v) instructions of the alignment of the content to the goal (what we call *format* in Table 1), vi) constraints and limitations[7]. The prompts were first produced in Italian and were built according to specific linguistic, lexicographic and methodological criteria, as well as to a set of parameters which were conceived as useful to the analysis; some prompts also included *gold data*, namely correct examples which have been validated and were used as training data. Table 1 summarizes the different features of each task, which have been adapted to our purposes. In bold we highlighted the keywords that have been used to induce the bot to generate the expected output (a sort of *killer tokens*, in NLP terms).

As Table 1 shows, even though instructions vary, all prompts ask for the identification of non-literal uses of BPNs and for a classification of the cognitive process involved (metaphor vs. metonymy). As for the output, all prompts aim at collecting a structured inventory (e.g. a table, a list, a semantic network) of authentic examples from explicit sources (e.g. dictionaries, newspapers, literary or web texts). The prompts mainly differ in terms of i) scope of the focus, which is more specific in prompts 1-2 and broader in prompts 3-6, and ii) presence of gold data to be used for training (absent in prompts 1-3).

As far as the content of the results is concerned, the bot consistently identifies expressions containing BPNs used metaphorically or metonymically, and the meanings assigned to them

Table 1 Prompt components.

	Task (instruction, focus, parameters, gold data)				Expected Output (conditions, content alignment and constraints on generated text)		
Prompt	**Instruction**	**Focus**	**Parameters for selection**	**Presence of gold data**	**Expected content**	**Format**	**Constraints**
1	Build a **lexicographic entry**	BPN ***testa*** 'head'	**All abstract domains** (e.g. HIERARCHY, INTELLECT)	no	**MWUs and idiomatic expressions** with metaphorical and metonymic **shifts**	Represent as a **network** of metaphorical/ metonymic shifts	examples identified by using **dictionaries** (GRADIT and Treccani) and extracted from the **web**
2	As Prompt 1	**All** BPNs	**Abstract domain** of AGENCY/ CONTROL	no	As Prompt 1	no	As Prompt 1
3	**Identify** occurrences in text	BPN ***mano*** 'hand' in the text *Il barone rampante* by Italo Calvino	**All abstract domains** (non literal)	no	List of occurrences of **abstract uses**	First 50 occurrences	Identify and analyse a specific text and produce a **list**
4	**Identify** metaphorical and	BPN ***testa*** 'head'	Sources: **Treccani** + **De Mauro**	the word **testa** 'head' has **metaphorical/ metonymic** shifts	**Metaphorical/ metonymic uses**	First 3 occurrences grouped by abstract domains	Examples from two sources

(*continued*)

Table 1 Continued

	Task (instruction, focus, parameters, gold data)				Expected Output (conditions, content alignment and constraints on generated text)		
Prompt	**Instruction**	**Focus**	**Parameters for selection**	**Presence of gold data**	**Expected content**	**Format**	**Constraints**
	metonymic uses						
5	As Prompt 4	As Prompt 4	As Prompt 4 + Corriere della sera and La Repubblica	as Prompt 4	At least **25 expressions**	Numbered table that highlights meaning, abstract domains, 3 **examples**	Examples from all 4 sources; indicate source
6	As Prompt 4	As Prompt 4	As Prompt 4 + La Repubblica	**Example of MWE**: "avere la testa fra le nuvole" **Example of sentence**:: "Marco ha sempre la testa fra le nuvole e non sta mai attento a quello che dico".	At least **25 examples**	Table with 4 columns: i. meaning of the expression, ii. abstract semantic domains, iii. 3 **sentences**	All 3 sentences per expression from *La Repubblica* with link

are accurate. Its effectiveness increases with the progression (and the complexification) of prompts, and in particular with the clearer description of the expected output: while in the results of prompt 1 it is possible to find some mistakes (e.g. *foto a testa in giù* 'upside-down photo'), this does not happen in the latter ones. When a high number of examples is required (e.g. prompt 3, 5, 6), the query has to be repeated, because the system does not return enough data in a single run, and always asks for a validation. One of the experiments that were repeated in the prompts was related to the use of the trigger word "usage example", which has caused some issues in the output. In the prompt, the term was intended to elicit authentic, contextualized examples. However, in the results of the first two prompts, the examples often appeared as decontextualized text chunks[8]. Consequently, we had to specify that usage examples should be provided as "sentences" extracted from actual/specific sources. As for the format, the bot does not create a lexicographic entry (as it was not properly trained for this purpose), but it provides a good table representation. By contrast, the semantic network representation often contains inaccuracies due to misclassification of metaphors and metonymies. This highlights one of the major limitations of the bot: while it is effective at extracting linguistic data (examples and authentic occurrences of BPNs) it is not always able to interpret cognitive processes or abstract semantic domains. Indeed, the classification of metaphor vs. metonymy is often unreliable (e.g. the expressions avere la testa tra le nuvole 'to be distracted' and capo di stato 'head of state' are wrongly interpreted as metonymies rather than metaphors)[9]. Furthermore, some domains are frequently repeated (different abstract concepts are assigned to the same expression, but they are usually synonymous rather than truly distinct) and others are closely tied to individual examples: they often lack the level of abstraction that a human mind can provide. In sum the system focuses more on retrieving expressions than differentiating semantic domains.

3.2. Approaching the best prompt: input structure and output evaluation

3.2.1 Input

Our final prompt is one in which we brought together all the requirements and observations collected during the initial experiments. This prompt was first applied to the BPN *testa* (Prompt 7_*testa*) and subsequently to the BPN *mano* (Prompt 8_*mano*).

The prompt is presented to ChatGPT in two steps. The first step includes an introduction to the topic, which presents and relates the key terms, as well as the actual instructions, which revolve around the data and the structure of the requested table. Here, the necessary features of the output are made as explicit as possible. The second step provides an additional instruction based on the results of the first step.

The level of granularity of this prompt is the result of several trials aimed at testing the system's effectiveness in relation to the number, type, and content of the information. For example, the creation of two successive steps stems from a "thematic" distinction that proved necessary to resolve interpretation issues in the response. Table 2 illustrates Prompt 7_*testa* along with its English translation.

In the following, we specify the components of the prompt, which underwent the most significant trigger adjustments:

- **Introduction:** here we replaced *literal sense* with *compositional sense*, following the observation of Phelps et al. 2024: "Interestingly, replacing the word 'Literal' with 'Compositional' did seem to have a positive effect"[10]; we also made more explicit the

Table 2 Prompt 7_*testa*

Prompt 7_testa	EN translation
STEP 1: La parola “testa” viene usata in italiano non solo in senso composizionale ma anche come metafora o metonimia. In questi casi assume un significato diverso, associato a un dominio astratto, ad esempio GERARCHIA o LEADERSHIP. Individua nel Vocabolario Treccani e nel Dizionario De Mauro (su L’Internazionale) 25 espressioni metaforiche e metonimiche con “testa” e compila una tabella con le seguenti colonne: 1. Espressione contenente “testa”, ad esempio “essere in testa”, “avere la testa fra le nuvole”, indicando se si tratta di metafora, metonimia (o altro) 2. Significato dell’espressione 3. Dominio astratto di riferimento (cerca di usare etichette più generiche, che possono adattarsi a diverse espressioni, come GERARCHIA SOCIALE, GERARCHIA SPAZIALE, LEADERSHIP, INTELLETTO) 4. Tre frasi che illustrino l’uso dell’espressione, ad esempio: “Il cd si trova in testa alla classifica”, oppure “Marco ha sempre la testa fra le nuvole e non sta mai attento a quello che dico”. Le frasi devono essere tratte dal quotidiano La Repubblica e/o Il Sole 24 Ore. Indica il link di ciascuna frase. Tutti e quattro i campi sono obbligatori. Numerare le righe della tabella. Produci subito la tabella completa di 25 righe, senza fasi intermedie. STEP 2: Sulla base della tabella precedente, prova a mettere in relazione tra loro i domini astratti in una rete concettuale che evidenzi i passaggi da un significato all’altro.	STEP 1: The word “testa” (‘head’) is used in Italian not only in a compositional sense but also as a metaphor or metonymy. In these cases, it takes on a different meaning, associated with an abstract domain, for example HIERARCHY or LEADERSHIP. Identify, in the Treccani Dictionary and in the De Mauro dictionary (on L’Internazionale) 25 metaphorical and metonymic expressions with “testa”, and compile a table with the following columns: 1. Expression containing “testa”, for example “essere in testa” (‘to be in the lead’), “avere la testa tra le nuvole” (‘to be distracted’) indicating whether it is a metaphor, metonymy (or other). 2. Meaning of the expression 3. Abstract domain (try to use generic tags that can adapt to different expressions, such as SOCIAL HIERARCHY, SPATIAL HIERARCHY, LEADERSHIP, INTELLECT. 4. Three sentences illustrating the usage of the expression, for example: “Il cd si trova in testa alla classifica” (‘The cd leads the hit parade’), or “Marco ha sempre la testa fra le nuvole e non sta mai attento a quello che dico” (‘Marco always has his head in the clouds and he never pays attention to what I say’). The sentences must be sourced from La Repubblica and/or Il Sole 24 Ore newspapers, with links for each sentence. All four fields are mandatory. Number the rows of the table. Produce the complete table with at least 25 rows, without intermediate steps. STEP 2: Based on the previous table, try to relate the abstract domains to each other in a conceptual network that highlights the shifts from one meaning to another.

distinction between the meaning of the expression (initially labelled *semantic value*) and the relevant abstract domain (initially labelled *abstract semantic value*).

- **Gold data:** the prompt contains gold data concerning idiomatic expressions, abstract domains, and usage examples.
- **Sources of idioms:** *Vocabolario Treccani* and *Dizionario De Mauro* (*Il Nuovo De Mauro*), while eliminating *La Repubblica* from these sources, since indicating non-lexicographic sources often led to unclear or incorrect results.
- **Sources of examples:** in this case we expanded the number of sources by adding, alongside *La Repubblica*, also the newspaper *Il Sole 24 Ore*, in an attempt to avoid otherwise existing gaps.
- **Specification of a precise number of expressions:** given the high productivity of BPNs in terms of idiomatic expressions, we considered 25 expressions to be a good average; the limitation of a high number of entries in the table is that the system sometimes reacts by providing the data not all at once but progressively, for instance in groups of three or five (this occurs even when explicitly asked to provide all the data in a single step).
- **Table with precise structural features** in terms of number and type of columns.

In the next section, the output of Prompt 6a_*testa* and Prompt 6b_*mano*[11] will be discussed and compared.

3.2.2 Output

To evaluate the output produced by the last prompt, the following different features are taken into account: i) the kind of expressions selected, ii) the presence of the requested authentic examples and relevant links, iii) the explanation of their meanings, iv) the correct identification of the cognitive process, v) the correct interpretation of the abstract domain, as well as vi) an effective representation of the semantic network of abstract domains and expressions.

Prompt 7 performs optimally at identifying several kinds of multiword units and idiomatic expressions (feature i). Indeed, different kinds of multiword units are included in the output, such as nominal (e.g. *testa di legno* 'forehead', *mano di ferro* 'iron fist'), verbal (e.g. *essere in testa* 'to be in the lead', *dare una mano* 'to give a hand), adjectival (e.g. *a testa bassa* 'head down', *a mano armata* 'armed') as well as adverbial (e.g. *testa a testa* 'head to head') multiwords. It is worth noting that most of them are verbal units (62% of the examples with *testa* and 88% of examples with *mano*). When present[12], corresponding authentic examples (feature ii) are representative of the uses of the unit; however, they do not always comply with lexicographic requirements for usage examples and should undergo at least partial revision[13]. The meaning associated with the unit (feature iii) is represented as a brief gloss and is always relevant[14]. In some cases, it also reveals semantic differences arising from polysemy; details concerning polysemy are taken from the dictionaries mentioned in the prompt. This is the case of the nominal multiword *testa di legno*, whose meaning can be either associated with 'an ignorant person' or 'a figurehead' (cf. *Il Nuovo De Mauro*). Interestingly, polysemy is not illustrated in authentic examples, where the different senses are not accurately represented[15]; for instance, for *testa di legno*, only the value of 'figurehead' is provided in the examples:

(1) Un prestanome, una testa di legno dietro cui si nascondeva il boss
'A front man, a figurehead behind whom the boss hid'
(2) Una testa di legno usata per operazioni fiscali

[14] In meaning definitions, our results were generally more encouraging than those of previous studies, such as Lew (2023) and Jakubíček & Rundell (2023).

[15] On this issue, see also Jakubíček & Rundell (2023), showing similar results.

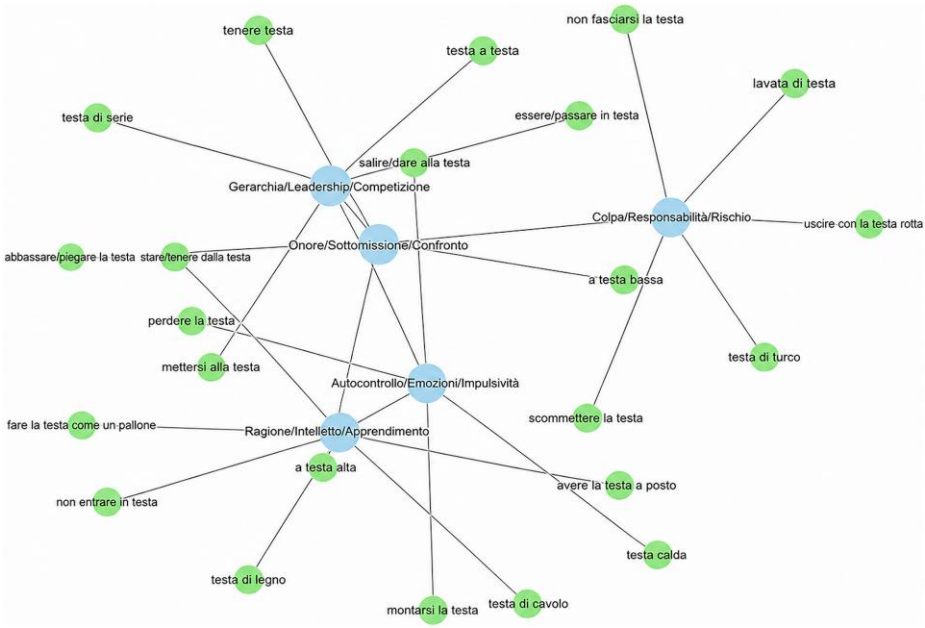

Figure 4 Semantic network of domains/expressions with the BPN *testa* 'head'.

'A figurehead used for financial operations'

Despite achieving excellent results in identifying sequences, the bot we used for the experiments does not always perform equally well in analysing them, and consequently in linking these expressions to the cognitive processes involved and to abstract domains. On the one hand, the system shows a reduced ability to recognize metonymies. Indeed, its metonymy identifications are often incorrect: as an example, this is the case of *testa di serie* 'top of the league' and *restare a mani vuote* 'to be left empty-handed', which are wrongly attributed to the process of metonymy. On the other hand, despite the exemplificative instruction, the tags selected for abstract domains are always closely tied to the semantics of the expression as a whole: in these cases, the bot is unable to generalize to a higher-level concept referred to the BPN and constantly focuses on the specific semantic nuances of the entire expression. For instance, *andare contromano* 'to go in the wrong direction' is associated with the tag concepts of 'violation and nonconformity' (i.e. the metaphorical meaning of the whole unit) instead of the spatial meaning of *mano*. This happens also for *non entrare in testa* 'to not understand/memorize', which is associated with the tag concept of 'learning' instead of the more general domain of INTELLECT. These are significant limitations of the system, which require a constant human manual post-editing. As a consequence, the semantic network created by the bot (feature vi) shows some weaknesses. The bot made an effort to group the values into as few domains as possible; however, beyond the issue related to the tags and higher-level concepts, it is worth noting an inconsistency between the expression itself and the broader domain to which it is assigned. Below the semantic network of *testa* is represented (Figure 4).

Summarizing, ChatGPT registers very good performances for tasks aiming at copying and extracting information from other sources. Its results vary considerably when any sort of analysis, beyond the generation task, is required. Table 3 displays the evaluation of the task output. Besides the evaluation of the overall performance for each feature, we also estimate the degree and type of

Table 3 Evaluation of task performance (Italian prompt).

Feature	Description	Task	Performance	Lexicographic PE
Feature i	Identification of expressions	Extraction	√	Validation and integration
Feature ii	Presence of authentic examples	Extraction	√	Adaptation
Feature iii	Explanation of meanings	Generation (by copying)	√	Validation and integration
Feature iv	Identification of the cognitive process	Generation + analysis	Not always adequate	Only advisable in specific cases
Feature v	Interpretation of the abstract domain	Generation + analysis	Unsuccessful	-
Feature vi	Semantic network of domains/ expressions	Generation + analysis	Unsuccessful	-

post-editing (PE) required to reach dictionary-like quality. For features iv-vi, post-editing could be excessively time-intensive and thus less suitable than a manual lexicographic approach.

According to our results, the system can perform mechanical tasks of extraction and reformulation that do not require interpretation by means of specific knowledge. However, it is unable to identify implicit relations between objects and to generalize rules and generate abstraction as a human would do. As shown in the table, this carries significant implications for the way the lexicographic process is planned.

3.3. Is a chatbot able to automatically generate valid prompts for our purposes?

The last challenge that is worth describing in this paragraph is the potential that the ChatGPT has in generating prompts. In fact, we asked the system to reformulate our best prompt (Prompt 6a_*testa*) in order to create a better one, suited to our needs. Its reformulation produced a prompt that was not highly schematic, which included, in the form of an outline:

(i) the actions to be carried out (e.g., identifying multiword expressions, using specific sources)
(ii) the parameters (the cognitive processes of metaphor and metonymy, with specific examples)
(iii) the trigger word (i.e. *testa*)
(iv) the criteria of inclusion (namely the kind of word combinations that have to be selected, such as idiomatic expression, multiword units or collocational patterns) and of exclusion (namely the literal uses of the BPN)
(v) the method of identification of the expressions
(vi) the output format (which schematically includes all our previous requests)
(vii) quality validations

Therefore, two major differences between our prompt and the one that was automatically produced by Chat GPT5 can be highlighted. On the one hand, a stylistic difference emerges:

Table 4 Excerpt from the automatically generated prompt.

Excerpt (IT)	EN translation
Dal Vocabolario Treccani e dal Dizionario De Mauro, individuare le voci e i sensi relativi a "testa" che corrispondono a usi figurati o metonimici; registrare il link alla voce.	From Treccani Vocabulary and De Mauro dictionary, identify the entries and the senses of "head" which correspond to figurative or metonymic uses; record the link to the entry.
Per ciascuna espressione, recuperare 3 frasi attestate da La Repubblica, all'interno dell'intervallo temporale 1990–oggi (modificare se necessario).	For each expression, retrieve 3 attested sentences from La Repubblica, in the temporal range 1990-today (modify if necessary).
Per ciascuna attestazione, fornire: data (AAAA-MM-GG), URL, e un breve estratto ≤90 caratteri (per rispettare il fair use e i diritti IP).	For each occurrence, provide: date, URL and a brief extract ≤90 characters (to respect the fair use and the IP rights).
Associare ciascuna espressione a un dominio astratto.	Associate each expression to an abstract domain.

the automatically generated prompt lacks discursive argumentation and is extremely schematic[16]. On the other hand, considerable differences emerge in terms of the content of the actions required. Precise instructions related to the method of identification of expressions are included in the prompt, even those that were not required in our first draft (e.g. the time range, date, URL and a brief extract ≤90 characters to respect the fair use and the IP rights). Table 4 provides a brief extract of the prompt:

The automatically generated prompt also includes instructions for quality verifications, aiming at avoiding repetitions and at ensuring a representative set of examples. The most interesting aspect is that, contrary to our expectations, the output produced by the automatically generated prompt performs significantly worse than ours. First of all, it proceeds through several intermediate steps, continuously asking for confirmation about the adequacy of the expected results and showing only a few results at a time. After several rounds of verification of data quality, the tag for the abstract domain is generally not appropriate, and the number of authentic examples obtained is still lower, as is their quality (it only shows the link to the examples, and in some cases the examples are actually taken from dictionaries). Therefore, going back to our previous question "Is a chatbot able to automatically generate valid prompts for our purposes?": ChatGPT seems to produce a more complex prompt which includes too many and too much specific instructions. Beyond a more sophisticated process of data extraction and data analysis, the automatically generated prompt produces an output with a lower level of quality.

3.4. Do chatbots perform better when working with English?

The last experiment we drove consisted in creating an English prompt equivalent to the Italian prompt 6a, to check its performance for a different language. Since English is the primary language on which ChatGPT is trained[17], a better output was expected in terms of both extraction

[16] As He et al. (2024) notice, the format of the prompt has significant consequences on the GPT-based model performance.

[17] Many LLMs are mainly trained on English data: as an example, as Mondshine et al. (2025: 1) note "GPT-3 was trained on 119 languages, but only 7% of the tokens are from non-English languages".

Table 5 Prompt 9_*head.*

Prompt 9_head
The word "head" in English is used not only in a compositional sense but also as a metaphor or metonymy. In these cases, it takes on a different meaning, associated with an abstract domain, such as HIERARCHY or LEADERSHIP. Identify, in the Cambridge English Dictionary online and/or in the Collins Online Dictionary, at least 15 metaphorical and metonymic expressions with "head", and compile a table with the following columns: Expression containing "head", for example "to keep the head about oneself", indicating whether it is a metaphor, metonymy, or other. Meaning of the expression Abstract domain of reference (try to use more general labels that can apply to different expressions, like SOCIAL HIERARCHY, SPATIAL HIERARCHY, LEADERSHIP, INTELLECT) Three sentences illustrating the usage of the expression, for example: "Though he was mad, he kept his head about him in public" The sentences must be sourced from The Guardian, The Independent, or other UK newspapers, with links for each sentence. All four fields are mandatory. Number the rows of the table. Produce the complete table with at least 15 rows, without intermediate steps.

and interpretation of data. To this purpose, the whole prompt was translated and specifically adapted for English: therefore, Italian expressions and examples were replaced with English counterparts, and English dictionaries and newspapers were selected for the extraction of authentic examples[18]. For the sake of clarity, the English prompt is represented below, in Table 5:

As for the results, it is possible to notice some similarities with Italian. First, there is a prevalence of verbal structures containing BPNs (66%); adverbs, nouns, and prepositional phrases are also found, but in a lower number. Unlike the results of Italian, however, those in English are characterized by the occurrence of similar expressions, which is marked as the presence of a "variant" (for example, *bang your head on the door* is repeated as *banging your head against the wall*, *bang your head against a (brick) wall*; and the sequence *get your head around (something)* is repeated as *get your head around end-of-year exams* and *couldn't quite get your head around*). Thus, the bot fails to distinguish between variation and mere repetition, and proves to be imprecise and redundant in data extraction. However, while the selection of English expressions does not register a good performance, the authentic examples extracted from newspapers and the relevant semantic glosses are adequate. Another common feature with Italian is represented by the identification of the cognitive process involved in the abstract use of the BPN: the recognition of metonymy, in particular, as for Italian, is often incorrect. For instance the nominal MW *big head* referring to 'someone who is conceited or arrogant' is tagged as "metonymy" and the sequence *from head to toe* is labelled as "literal meaning" instead of "spatial metaphor". English results differ from those of Italian with regard to the labelling of abstract domains. In particular, English tags are usually more pertinent. For example, the tag TOTALITY is applied to the sequence *from head to toe*, while the tag INTELLECT is used for all sequences in which the *head* is referred to as a container of ideas. Last, the semantic-network representation contains some inaccuracies referring to the grouping of different abstract domains, and therefore to their relationships: for instance, it is not clear how the domain of TOTALITY could relate to that of INTELLECT that is directly linked to it. Thus, despite a more accurate domain labelling, the overall English performance in the creation of semantic networks remains rather limited. Figure 5 represents the semantic network of *head*:

[18] In particular, the choice of dictionaries and newspapers mostly depends on website limitations: many of the English dictionaries/newspapers that are freely accessible online, only allow a limited search, because of paywalls.

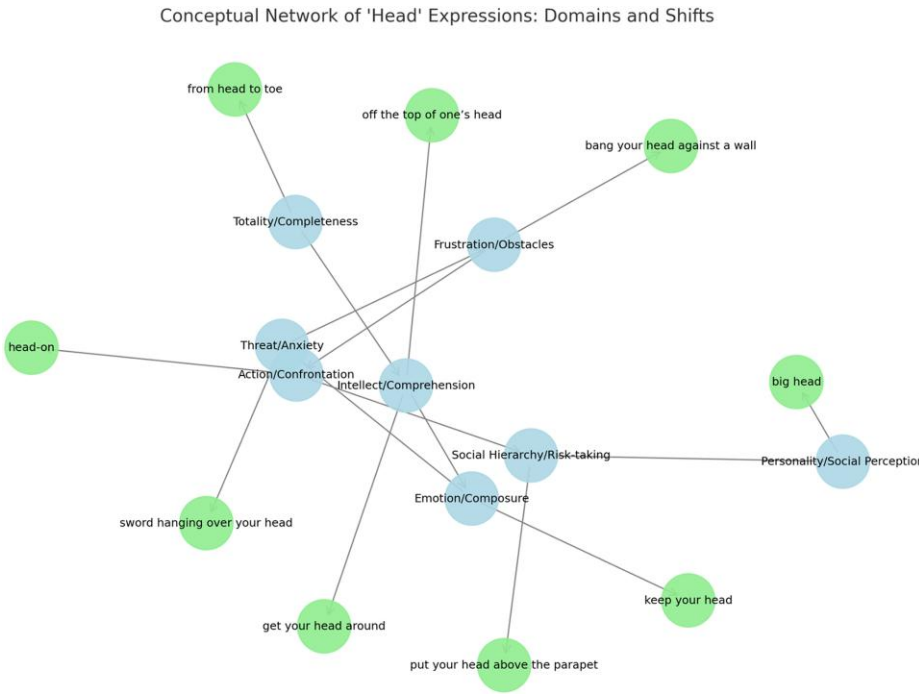

Figure 5 Semantic network of domains/expressions with the BPN *head*.

Table 6 Evaluation of task performance (English prompt).

Feature	Description	Task	Performance	
			IT	EN
Feature i	Identification of expressions	Extraction	√	Not adequate
Feature ii	Presence of authentic examples	Extraction	√	√
Feature iii	Explanation of meanings	Generation (by copying)	√	√
Feature iv	Identification of the cognitive process	Generation + analysis	Not always adequate	Not adequate
Feature v	Interpretation of the abstract domain	Generation + analysis	Unsuccessful	√
Feature vi	Semantic network of domains/expressions	Generation + analysis	Unsuccessful	Not always adequate

Summarizing, the performances for English are not particularly strong. This implies that even for the language on which LLMs are most extensively trained, a post-editing phase for the evaluation and integration of the data is necessary. Table 6 compares the English performance with the Italian one (cf. Table 3):

The grey cells highlight the critical areas, where ChatGPT performs worse. Despite expectations, data extraction for English is rather problematic: this means that, alongside the more challenging task of analytic generation, extractive tasks may also underperform even in high-resource languages[19]. This leads us to conclude that the mere quantity of training data available to a model is not sufficient to make it intelligent enough to carry out a high-performance data extraction/analysis.

4. Conclusions

This study presents a series of exploratory experiments on the use of generative AI for the processing of idiomatic expressions related to BPNs for lexicographic purposes. The innovative contribution of our approach lies in the focus both in the acquisition of such expressions and in the elaboration of the way in which the resulting data can be presented within a general monolingual dictionary or a learner's dictionary, namely by grouping them along abstract domains[20]. The conceptual domains presented in this study should not be understood as absolute or exhaustive classifications, but rather as indicators which, much like sense disambiguators or "signposts" used in other dictionaries, may help users classify idiomatic expressions and highlight similarities among them. At the same time, we do not regard these indicators as arbitrary; rather, we believe that they may be shared by a large number of users, although this assumption would of course need to be assessed through empirical testing. The domains presented in this study are those most extensively discussed in the relevant literature. The assignment of each expression to a particular domain is based on the semantic nuance that most closely reflects its meaning. The domains themselves are interrelated and overlap conceptually, rather than constituting separate and mutually independent categories.

The overall results obtained from the experiments show the effectiveness (as well as the drawbacks) of this method, confirming some of results identified in earlier studies (e.g. Lew, 2023; Jakubíček & Rundell, 2023).

As initially stated, in the lexicographic process, retrieving relevant information from scratch—starting from the idiomatic expressions themselves and their possible organization—is extremely time-consuming and requires the combination of different empirical methods (e.g., corpus linguistics, consultation of existing dictionaries, introspection). The experience with the *Phrase-based Active Dictionary* highlights that, for opaque and complex phraseological data such as idiomatic expressions, a new methodology is needed, one that can be integrated into a hybrid framework combining:

- a supervised automated phase, involving the elaboration of prompts that approximate the desired result as closely as possible;
- a post-editing phase, in which the lexicographer's main task is to validate, enrich, and adapt the output previously obtained;
- a possible further phase of fine-tuning of the prompts, especially when applying the model to new data (e.g., moving from BPNs to a different semantic field).

[19] On this issue, see also Mondshine et al. (2025, p. 2), who demonstrate that "in extractive tasks […], where the output overlaps with the provided context and no generation is needed, the model is either agnostic to the context language in the case of high-resource languages or prefers context in the source language in the case of low-resource languages". As regards the level of recall reached in the present study, it can be stated that the system did not identify all possible uses (the authors had, however, requested only a limited number of examples), nor did it consistently retrieve the most frequent ones.

[20] As the anonymous reviewer notices, it is worth noting that domain-based grouping may not apply to all idiomatic expressions; this may depend on the individual history of each expression.

The use of generative AI represents an expansion, not a replacement, of other methods, which remain central in the post-editing stage.

Our general observations regarding the elaboration and fine-tuning of prompts can be summarized as follows:

(a) Unstable effects of fine-tuning. Adjusting prompts does not necessarily yield a consistent improvement in results. Parameters (triggers) that initially produced satisfactory outputs may, in subsequent iterations, lead to poorer performance even without modification.
(b) Importance of gold data. Supplying gold data is crucial for the system's correct interpretation and execution of the prompt.
(c) Distinction between extraction tasks and generation tasks: prompts need to account for both modes, and the output should be evaluated separately.
(d) Persistence of trial-and-error prompting. Even when a prompting framework with predefined patterns is in place (cf. White et al., 2023), the process remains largely reliant on trial and error. Our objective has been to approximate, as closely as possible, a reusable model tailored to the lexicographic processing of idioms, taking into account a subsequent post-editing phase.

We acknowledge that relying solely on ChatGPT-5 Plus in this case study requires us to interpret our results with caution. Still, our findings underscore the strong potential of applying generative AI to the study of idiomatic expressions related to BPNs. Future work should test the performance of further LLMs, expand this approach to broader areas of the lexicon and investigate its applicability in contrastive settings, with the aim of advancing AI-assisted methodologies for lexicographic practice and of improving our understanding of their potential, as well as of their limitations.

Notes

1 We would like to thank the anonymous reviewer who provided valuable comments on the first version of the text. Although the paper is the fruit of a joint reflection and cooperation, for academic reasons Laura Giacomini is responsible for Sections 2, 2.1, 2.2, 3, and Valentina Piunno for Sections 1, 1.1, 3.1, 3.2, 3.3, 3.4. Section 4 is attributed to both authors.
2 "Speaking of language is part of an activity, or a form of life" (Wittgenstein [1953]1958: 11).
3 Cf. Evans and Green (2006) for a comprehensive overview of the various factors that contribute to meaning extensions and the main theoretical approaches addressing this issue.
4 As noted by an anonymous reviewer, assigning an idiom to an abstract semantic domain thus creating clusters of idioms may be problematic and not necessarily facilitate user access. The authors acknowledge this concern; however, they emphasize that grouping idioms into semantic domains constitutes an added value for the dictionary and represents an innovative features within the current lexicographic landscape.
5 It has been shown that the performance of the prompt may vary depending on the model used (Dang et al., 2022). For the sake of potential replicability of the analysis, we specify that the version employed in this study is the GPT-5 (Plus version).
6 The zero shot model has no access to training data and provides good results in simple tasks. The few-shot model is useful when the task is complex (it is composed of more subtasks) and when the output needs to be tailored according to specific purposes;

furthermore, the results of the few-shot task may vary depending on the model used (Phelps et al., 2024). According to Mondshine et al. (2025: 6), "the top-performing configurations of the GPT model […] show that the optimal configurations are those that include examples, i.e., a few-shot rather than zero-shot". Cf. also Liu et al. (2021) and references therein included for a detailed analysis of the two learning models.

7 While the former three refer to the task itself (the *instruction*, the *context* and the *input data* in the terms of Giray, 2023), the latter focus on the expected output.

8 On this issue, see also Lew (2023) and Jakubiček & Rundell (2023).

9 This may be due to the fact that, in MWEs including BPNs the whole sequence (and not just the BPN) may undergo a (set of different) semantic shifts. This brings to a lexicalized unit with a non-compositional meaning.

10 In our case, however, this did not produce any significant effect.

11 For Prompt 8_*mano*, we indicated CONTROL, POSSESSION and ACTION as potential abstract domains. We exemplified the expressions with the sequences *sfuggire di mano* 'get out of hand' and *per mano di* 'by the hands of'. As authentic examples we provided *A Marco è sfuggita di mano questa situazione* 'Marco got out of hands this situation' and *Il ladro è stato arrestato per mano della polizia* 'The thief was arrested by the hands of the police'.

12 It is important to note that they are not always represented in the results.

13 This confirms what was highlighted in earlier studies by Lew (2023) and Jakubíček & Rundell (2023).

14 In meaning definitions, our results were generally more encouraging than those of previous studies, such as Lew (2023) and Jakubiček & Rundell (2023).

15 On this issue, see also Jakubiček & Rundell (2023), showing similar results.

16 As He et al. (2024) notice, the format of the prompt has significant consequences on the GPT-based model performance.

17 Many LLMs are mainly trained on English data: as an example, as Mondshine et al. (2025: 1) note "GPT-3 was trained on 119 languages, but only 7% of the tokens are from non-English languages".

18 In particular, the choice of dictionaries and newspapers mostly depends on website limitations: many of the English dictionaries/newspapers that are freely accessible online, only allow a limited search, because of paywalls.

19 On this issue, see also Mondshine et al. (2025, p. 2), who demonstrate that "in extractive tasks […], where the output overlaps with the provided context and no generation is needed, the model is either agnostic to the context language in the case of high-resource languages or prefers context in the source language in the case of low-resource languages". As regards the level of recall reached in the present study, it can be stated that the system did not identify all possible uses (the authors had, however, requested only a limited number of examples), nor did it consistently retrieve the most frequent ones.

20 As the anonymous reviewer notices, it is worth noting that domain-based grouping may not apply to all idiomatic expressions; this may depend on the individual history of each expression.

References

A. Dictionaries

Bosque Muñoz, I. (2004). *REDES: Diccionario combinatorio del español contemporáneo las palabras en su contexto*. Ediciones SM.

[CLD] Cambridge Learner's Dictionary: Definitions & Meanings. (1999). Available from https://dictionary.cambridge.org/dictionary/learner-english/

[COD] Collins Online Dictionary. (2007). Available from https://www.collinsdictionary.com/

Il Nuovo De Mauro. Available from https://dizionario.internazionale.it

[LDOCE] Longman Dictionary of Contemporary English. (2008). Available from https://www.ldoceonline.com/

[OALD] Oxford Advanced Learner's Dictionary. (2020). Available from https://www.oxfordlearnersdictionaries.com/

Schafroth, E., Bürgel, M., Imperiale, R., & Martulli, F. (2018). *GEPHRI (Gebrauchsbasierte Phraseologie des Italienischen)*. Romanistik IV. University of Düsseldorf. Available from https://gephri.phil.hhu.de/

B. Other literature

Arias-Arias, I., Vázquez, M. J. D., & Riveiro, C. V. (2024). Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: kann ChatGPT die lexikographische Textsorte nachbilden? *Lexikos*, *34*, 51–76. https://doi.org/10.5788/34-1-1879

Balbachan, F. (2006). Killing time: Metaphors and their implications in lexicon and grammar. In H. Clarenz-Löhnert, M. Döring, K. Gabriel, K. Mutz, D. Osthus, C. Polzin-Haumann, & N. Roßbach (Eds.), *Metaphorik.de 10* (pp. 6–30). https://doi.org/10.17185/duepublico/83541

Brown, C. H. (1976). General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature. *American Ethnologist*, *3*(3), 400–424. http://www.jstor.org/stable/643763

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). *How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models.* arXiv:2209.01390. https://doi.org/10.48550/arXiv.2209.01390, preprint: not peer reviewed.

de Schryver, G.-M. (2023). Generative AI and Lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387. https://doi.org/10.1093/ijl/ecad021.

DiMuccio-Failla, P., & Giacomini, L. (2022). A proposed microstructure for a new kind of active learner's dictionary. *Lexicographica*, *38*(1), 475–499.

DiMuccio-Failla, P. (2025). A theory for a usage-based cognitive lexicography. In L. Giacomini, & V. Piunno (Eds.), *Patterns of meaning in lexicography and lexicology* (pp.1–14). Berlin: De Gruyter.

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching & Learning Practice*, *20*(5). https://doi.org/10.53761/1.20.5.02

Enfield, N. J., Majid, A., & van Staden, M. (2006). Cross-linguistic categorisation of the body: Introduction, *Language Sciences*, *28*(2–3), 137–147.

Evans, N. (2010). Semantic typology. In Sung, J.J. (Ed.), *The Oxford handbook of typology* (pp. 504–533). Oxford University Press.

Evans, V., & Green, M. (2006). *Cognitive Linguistics. An introduction*. Edinburgh University Press.

Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st century skill. *Frontiers in Education*, 9. https://doi.org/10.3389/feduc.2024.1366434

Ganfi, V., Piunno, V., & Mereu, L. (2023). Body part metaphors in phraseological expressions: A comparative survey of Italian, Spanish, French and English. *Languages in Contrast*, *23*(1), 1–33. https://doi.org/10.1075/lic.21006.gan

Giacomini, L. (2025). Introduction to the PhraseBase project. In L. Giacomini, & V. Piunno (Eds.), *Patterns of meaning in lexicography and lexicology* (pp.15–18). Berlin: De Gruyter.

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*, 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Goddard, C. (2002). Whorf meets Wierzbicka: Variation and universals in language and thinking. *Language Sciences*, *25*(4), 393–432. https://doi.org/10.1016/S0388-0001(03)00002-0

Hanks, P., & Ježek, E. (2008). *Shimmering lexical sets*. In E. Bernal, & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 391–401). Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Hausmann, F. J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts*, *31*(4), 395–406.

He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan S. (2024). *Does prompt formatting have any impact on LLM performance?* arXiv:2411.10541. https://doi.org/10.48550/arXiv.2411.10541, preprint: not peer reviewed.

Heine, B. (1997). *Possession: Cognitive sources, forces, and grammaticalization*. Cambridge University Press.

Heine, B., & Kouteva, T. (2002). *World lexicon of grammaticalization*. Cambridge University Press.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization* (2nd edn). Cambridge University Press. https://doi.org/10.1017/CBO9781139165525

Jakubiček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medved (Ed.), *Proceedings of the eLex 2023 Conference* pp. (504–518). Brno, Czech Republic.

Ježek, E. (2023). 7 Semantic co-composition in light verb constructions. In A. Pompei, L. Mereu, & V. Piunno (Eds), *Light verb constructions as complex verbs: features, typology and function* (pp. 221–238). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110747997-008

Johnson, M. L. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. The University of Chicago Press.

Koptjevskaja-Tamm, M., Vanhove, M., & Koch, P. (2007). Typological approaches to lexical semantics. *Linguistic Typology*, *11*(1), 159–185. https://doi.org/10.1515/LINGTY.2007.013

Kövecses, Z. (2000). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the Flesh: The embodied mind & its challenge to western thought*. Basic Books.

Lakoff, G. & Johnson, W. (1980). *Metaphors we live by*. University of Chicago Press.

Lehmann, C. ([1982]1995). *Thoughts on grammaticalization*. LINCOM Europa.

Lehmann, C. (2022). Foundations of body-part grammar. In R. Zariquiey, & P. M. Valenzuela (Eds.), *The grammar of body-part expressions* (pp. 14–76). Oxford University Press. https://doi.org/10.1093/oso/9780198852476.003.0002

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, *10*(1), 1–10. https://doi.org/10.1057/s41599-023-02119-6

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*(1), 1–8. https://doi.org/10.1057/s41599-024-02889-7

Lipka, L. (1992). *An outline of English Lexicology: Lexical structure, word semantics and word-formation*. Max Niemeyer Verlag Tübingen.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing*. arXiv:2107.13586. https://arxiv.org/pdf/2107.13586, preprint: not peer reviewed.

Majid, A., Enfield, N. J., & van Staden, M. (2006). Parts of the body: Cross-linguistic categorisation. *Special issue of Language Sciences*, *28*(4), 137–360.

Manerko, L. (2014). From human body parts to the embodiment of spatial conceptualization in English idioms. In G. Rundblad, A. Tytus, O. Knapton, & C. Tang (Eds.), *Selected Papers from the 4th UK Cognitive Linguistics Conference* (pp. 195–213). London: UK-CLA.

McKean, E., & Fitzgerald, M. (2024). The ROI of AI in lexicography. *Lexicography. Journal of ASIALEX*, *11*(1), 7–27. https://doi.org/10.1558/lexi.27569

Michaelis, L. A. (1993). 'Continuity' within three scalar models: The polysemy of adverbial still. *Journal of Semantics*, *10*, 193–237.

Michaelis, L. A. (1996). Cross-world continuity, and the polysemy of adverbial still. In G. Fauconnier, & E. Sweetser (Eds.), *Space, worlds and grammar* pp. (179–226). Chicago: The University of Chicago Press.

Mondshine, I., Paz-Argaman, T., & Tsarfaty, R. (2025). *Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs*, arXiv:2502.09331. https://doi.org/10.48550/arXiv.2502.09331, preprint: not peer reviewed.

Phelps, D., Pickard, T., Mi, M., Gow-Smith, E., & Villavicencio, A. (2024). *Sign of the times: Evaluating the use of large language models for idiomaticity detection*. arXiv:2405.09279. https://doi.org/10.48550/arXiv.2405.09279, preprint: not peer reviewed.

Piunno, V. (2023). The metaphorical/metonymic values of hand in texts of 1400-1600s. *Status Quaestionis*, *25*, 147–176.

Ptasznik, B., Wolfer, S., & Lew, R. (2024). A learners' dictionary versus ChatGPT in receptive and productive lexical tasks. *International Journal of Lexicography*, *37*(3), 322–336. https://doi.org/10.1093/ijl/ecae011

Robert, S. (2008). *Words and their meanings: Principles of variation and stabilization*. In M. Vanhove (Ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations* (pp. 55–92). John Benjamins.

Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? *Asialex 2023 Proceedings* (pp. 9–17). https://asialex.org/pdf/Asialex-Proceedings-2023.pdf

Ruthrof, H. (2000). *The body in language*. The Cassell Press.

Short, C. E., & Short, J. C. (2023). The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*, *19*, e00388. https://doi.org/10.1016/j.jbvi.2023.e00388

Talmy, L. (2000). *Toward a cognitive semantics*. MIT Press.

White, T., Lin, C., Rees, J., Godefroy, A., & Arnold, D. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv:2302.11382. https://doi.org/10.48550/arXiv.2302.11382, preprint: not peer reviewed.

Wierzbicka, A. (1972). *Semantic primitives*. Athenäum-Verlag.

Wierzbicka, A. (2007). *Understanding cultures through their key words: English, Russian, Polish, German, and Japanese*. Oxford University Press.

Wierzbicka, A. (2014). *Imprisoned in English: The hazards of English as a default language*. Oxford University Press.

Wittgenstein, L. ([1953]1958). *Philosophical investigations*. Blackwell.