
# UNIVERSITÀ DI PISA

## DIPARTIMENTO DI FILOLOGIA, LETTERATURA E LINGUISTICA

**Corso di Laurea Magistrale in Linguistica**


# Casting Light on Idiom Flexibility: A Corpus-based Approach


**CANDIDATO**

Marco Silvio Giuseppe Senaldi

**RELATORE**

Chiar.mo Prof. Alessandro Lenci

**CORRELATRICE**

Chiar.ma Prof.ssa Giovanna Marotta

**ESPERTI ESTERNI**

Chiar.mo Prof. Pier Marco Bertinetto

Dott. Gianluca E. Lebani



**ANNO ACCADEMICO 2014/2015**

# TABLE OF CONTENTS

# ABSTRACT

The goal of this work is to assess the cognitive plausibility of corpus-based measures that capture the formal flexibility and the semantic idiosyncrasy of a sample of Italian idiomatic expressions. The 87 idioms in our dataset are taken from the study of Tabossi and colleagues (2011), who elicited normative judgments on 245 Italian idioms from 740 native subjects. We use *Shannon Entropy* (Shannon 1948) to measure the lexical and morphosyntactic variability of our expressions and *Distributional Semantic Models* (*DSMs*) (Lenci 2008; Turney & Pantel 2010) to represent their semantics. Our dataset is extracted from the *La Repubblica* corpus (Baroni et al. 2004) via *SYMPAThy* (*Syntactically Marked PATterns*) (Lenci et al. 2014; 2015), a format of data representation that encompasses both PoS-related and syntactic information to derive word combinations from corpora. Performing a series of stepwise multiple regression analyses, we find out that psycholinguistic judgments on idiom predictability, literality and syntactic flexibility can be modeled by an array of computational measures, composed of our entropic and distributional values, token frequency and the number of fully lexicalized arguments exhibited by each idiom.

This thesis is organized as follows. In *Chapter 1* we illustrate the concepts of *idiomaticity* (Cacciari & Glucksberg 1991; Nunberg et al. 1994) and *multiword expressions* (*MWEs*) (Sag et al. 2001; Masini 2012), reviewing the major theoretical, psycholinguistic and computational studies that have been conducted on the subject. In *Chapter 2* we give a definition of *word combinations* and describe the *constructionist* framework (Fillmore et al. 1988; Goldberg 1995; 2006; Croft 2003; Croft & Cruse 2004; Hoffmann & Trousdale 2013) we have adopted in our work. We then survey both pros and cons of PoS-based and syntax-based methods for the extraction of word combinations from corpora and present *SYMPAThy* (*Syntactically Marked PATterns*), a format of data representation that combines both the approaches (Lenci et al. 2014; 2015). In *Chapter 3* we describe the entropic indices and the distributional measures we have exploited. *Chapter 4* begins with a brief description of the normative data collected by Tabossi and colleagues (2011) from which we took the idioms in our dataset. We then report the description of our first experiments, including data extraction, the calculation of our corpus-based indices and the execution of the stepwise multiple regression analyses with Tabossi et al.'s rankings as dependent variables. We then report the second experiment, wherein a syntactic acceptability test on Italian idiomatic expressions was prepared and submitted via CrowdFlower (http://www.crowdflower.com). The resulting ratings are then compared with those previously elicited by Tabossi et al. (2011) and used as dependent variables in a second series of stepwise regressions with our corpus indices as predictors. We finally provide some *Conclusions* and suggest future directions of research.

# CHAPTER 1
# MULTIWORD EXPRESSIONS AND IDIOMS: A THEORETICAL, PSYCHOLINGUISTIC AND COMPUTATIONAL OVERVIEW

## 1.1. On the pervasiveness of Multiword Expressions in language

One of the core features of human language that have been highlighted the most in the past two centuries of linguistic thought is its *creativity*, commonly associated with Wilhelm von Humboldt's motto that language makes *"infinite use of finite means"* (*Unendlicher Gebrauch von endlichen Mitteln*; von Humboldt 1988 [1836]: 91). In humboldtian perspective, language plays a major role in the constitution of thought, which is in principle endless. More than one hundred years later, Chomsky (1965; 1966) construes such a statement as a forerunner of the basic generative tenets, that regard human beings as innately endowed with a finite set of rules permitting, through their recursive application, the generation and understanding of a potentially open-ended set of sentences. Language users can therefore understand or produce novel utterances they have never encountered before and in a *stimulus-independent* fashion, that is, they can unpredictably utter any kind of sentence in any context depending on their state of mind (Chomsky 1959). The notion of unboundedness and stimulus-independence in linguistic behavior dates back to Descartes (1649/1927: 360) that, conceiving it as the true discriminating factor between humans and other animals and machines, describes it as follows:

> "without any finite limits, influenced but not determined by internal state, appropriate to situations but not caused by them, coherent and evoking thoughts that the hearer might have expressed, and so on". (quoted in Chomsky 2000 : 17)

Properly speaking, it should be noticed that, in recalling von Humboldt's quote, Chomsky seems to have misread its original meaning: what is infinite about generative grammar is the set of sentences that are produced and not the domain of thought that language expresses, as von Humboldt actually intends (Weydt 1972). Nonetheless, it is this very notion of sentence creativity that we are more interested in. In Standard Theory (Chomsky 1957; 1965), a sentence is generated via phrase-structure rewriting rules. Consider Chomsky's example *The man hit the ball*: starting from the sentence symbol *S*,

non-terminal symbols are gradually replaced with other non-terminal symbols according to the rules of the grammar, until terminal symbols like *N* (for nouns) and *V* (for verbs) are expanded by single words. This derivation can be depicted by a tree graph:

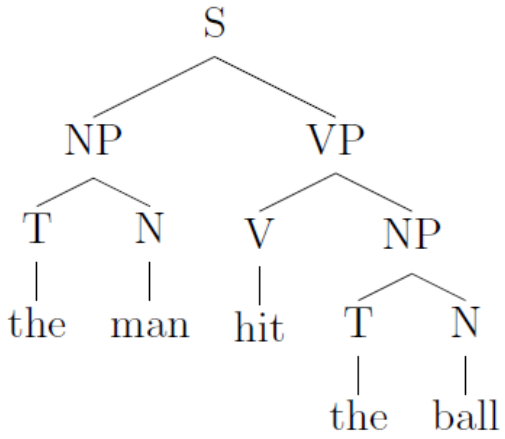



**Figure 1:** an example of sentence derivation (Chomsky 1957: 27)

Extended Standard Theory (Chomsky 1965) rethinks this mechanism by positing a *lexicon* and a set of *insertion rules* that position lexical items into the deep structure of a sentence. Surface structure is then derived by means of transformation rules (Chomsky 1965: 128 ff.). Crucially for the discussion at hand, such a view conceives the lexicon, a sort of repository for all those idiosyncrasies that cannot be generated by rules, as containing almost only single words and morphemes and being hence *"devoid of the combinatorial structure seen in phrases and sentences"* (Jackendoff & Pinker 2005: 219). The grammar is then responsible for taking these lexical entries and assembling them in phrases and sentences by rules that build, move and combine syntactic trees (ibid.).

All these stances, the existence of a clear-cut division of labor between lexicon and grammar, the conception of lexicon as an unordered list of single words and the idea that *"virtually every sentence that a person utters or understands is a brand-new combination of words"* (Pinker 1995: 22) have been addressed and challenged by a variety of theoretical frameworks and empirical findings in the following decades of linguistic research.

Prior to significant evidence coming from corpus linguistics (see below), the Chomskian notion that language is entirely novel in every production and comprehension act has been challenged by just two major counterexamples. First of all, this model does not account for the processing of idioms like *kick the bucket*, *pig in a poke* and *by and large*, which must be treated as *"ready-made surface structures"* (Watkins 1992: 392) having a direct link

between their phonological or graphemic form and their meaning to preserve their idiomatic interpretation (Chafe 1968; Lyons 1968: 177 ff.; Weinreich 1969; Jackendoff 1997). Anyway, since idioms constitute just a restricted list of expressions, Mainstream Generative Grammar has found a way out treating them as marginal exceptions that must be stored as a whole in the lexicon. In the second place, Pawley and Syder (1983: 193) observe that only a limited set of all the virtually possible sentences in a language are actually considered acceptable, ordinary and natural by the speakers of a language, while variant expressions with the same meaning but a different structure are labeled as "unidiomatic", "odd" and "foreignisms". In English, it is natural, when meeting someone, to wish them a *good morning* and not a *pleasant, fine* or *enjoyable morning* (Siyanova-Chanturia & Martinez 2014) or to describe a tea as *strong* and not *powerful,* although the meaning would be practically the same. Despite this, Chomskian grammar has avoided to focus on the practice of real speakers for a long time, given that it *"purports to be a description of the ideal speaker-hearer's intrinsic competence"* (Chomsky 1965: 4). Since it would not be incorrect to assert that a native speaker is perfectly able to build and decipher a sentence like *"The captain has illuminated the seatbelt sign as an indication that landing is imminent",* there would be no point observing that an actual speaker would rather utter something like *"The captain has put the seatbelt sign on, which means we're about to land"* (Wray 2002: 13). Until large corpora that could demonstrate the actual spread of pre-constructed linguistic sequences were available, idiomaticity and formularity have been mainly relegated to the field of sociolinguistics and pragmatics (Wray 2002: ibid.).

With the advent of corpus-based studies, extensive surveys have finally confirmed that an integral part of our spoken and written production is actually composed of *prefabricated* and *formulaic units*, rather than word-by-word assembled (Sorhus 1977; Sinclair 1991; Howarth 1998; Biber et al. 1999; Erman & Warren 2000; Wray 2002; Van Lancker-Sidtis & Rallon 2004; Siyanova-Chanturia & Martinez 2014).

Erman and Warren (2000) extract nineteen passages of 600 to 800 words from *The London Lund Corpus of Spoken English* (*LLC*) and the *Lancaster-Oslo-Bergen* corpus, representative for written English, and investigate how many slots (i.e. positions for a word) in a text like the following are filled by words representing single lexical choices and how many are parts of prefabricated expressions (*prefabs* in their terminology).

To the best of my knowledge, there is no record of a society which has used literacy for the

> profane and imaginative purposes and which has not produced books dealing with sexual topics. [G 77 001-004, Lancaster-Oslo-Bergen]

Just to foretaste the phenomenon we are going to illustrate in more detail, considering that the authors have underlined each lexical choice with a solid line, we can observe that the paragraph above contains 23 choices out of 33 slots:

> To the best of + my + knowledge, there is no record of a society which has used literacy for + the profane and imaginative + purposes and which has + not + produced books dealing with sexual topics.

The requisite for labeling a given combination of words as a prefab is *restricted exchangeability*, which means that at least one member of the expression cannot be substituted by a synonym without changing the meaning or the function of the whole. For instance, in saying *"They are good friends"*, we cannot replace *good friends* with *nice friends* without losing the idea of reciprocity; when using *I'm afraid* with the pragmatic function of softening bad news, we cannot change *afraid* into *scared* or *frightened*. Restricted syntactic variability is also used as a clue: expressions like *It will do* and the epistemological *I guess* cannot appear in another tense and be negated, respectively (e.g. **It does*, **I don't guess*). According to their meaning and function, Erman and Warren tell apart three categories of prefabs. Expressions like *good friends* or *to the best of one's knowledge* are named *lexical prefabs*, quantifiers (*a few*), links (*instead of*), introductors (*there is/are*), temporal and aspectual markers (*be going to*, *used to*) and so forth are called *grammatical prefabs*, while typical examples of *pragmatic prefabs* are discourse markers (*and then*, *I guess*) or performative routines (*thank you*, *good evening*).

Following Pawley and Syder (1983), Erman and Warren (2000) motivate the existence of such fixed chunks with the reflection that, in a given culture, it is natural to denote standard situations and to express oneself in typical social interactions by means of standard linguistic phenomena. To say it in Nattinger and DeCarrico's (1994) terms, *"just as we are creatures of habit in other aspects of our behavior, so apparently are we in the ways we come to use language".*

Interestingly, some prefabs have *open slots* that can be occupied by a more or less restricted set of words: in analyzing *to the best of my knowledge* in the given extract, a 'plus' sign is inserted to indicate that the fixed part is *to the best of … knowledge*, while the slot before *knowledge* must be filled by any sort of possessive element for the prefab to

be complete. Aside from the presence of open slots, other axes of variability are observed for such combinations. First of all, their average length appears to span from two to five words. Moreover, variation at the level of inflection and order is registered: a word-combination can occur with different determiners (*lay a/the table*), tense (*sit/sat down*) and voice (*the table is laid*), can be negated or modified by adverbs (*has not produced* in the extract above) and can exhibit variation in the reciprocal order of the elements (*it is going to / is it going to*). In any case, not every kind of modification is possible (*have a go at something, have another go at something* vs. **have the go at something*) but these restrictions are often unpredictable.

To sum up, what comes to the fore in this study is that 58.6% of the spoken texts and 52.3% of the written texts analyzed are composed of prefabricated expressions. These results have been corroborated by a great deal of evidence deriving from studies on written and spoken corpora. Sorhus (1977) finds 20% of formulaic expressions in a Canadian sample of spontaneous speech; Strässler (1982) likewise detects one idiom every four minutes and a half of discourse in conversational data of more than 100.000 words; Altenberg (1991; 1998) uses computer-search criteria to estimate that 80% of the London-Lund Corpus is represented by recurrent word-combinations; Biber et al. (1999) report that multi-word units constitute 28% of the spoken section and 20% of the written section of *Longman Spoken and Written English* corpus; according to Van Lancker-Sidtis and Rallon's (2004) analysis of the screenplay *Some Like It Hot*, nearly one fourth of the phrases and sentences uttered are speech formulas, idioms and proverbs. In discussing and revisiting the traditional generative assumptions on the lexicon, Jackendoff (1995) bases his argument that *"the theory of fixed expressions is more or less coextensive with the theory of words"* (Jackendoff 1995: 149) on evidence collected from the television show *The Wheel of Fortune*. Asking his daughter Beth to take note of all the phrases the contestants had to guess over a few months, he counts 10% of the whole corpus made of single words, 30% of compounds (*black and white film*, *Mexican peso*, *peanut butter*), 10% of idioms (*eat humble pie*, *I cried my eyes out*, *hit the road*), 10% of names (*John F. Kennedy*, *Addis Abeba*), 10% of meaningful names (*Democratic Convention*, *The Big Apple*), 15% of clichés (*any friend of yours is a friend of mine*, *gimme a break*, *time will tell*) and 5% of titles (*All You Need Is Love*, *Good Morning America*). All these examples represent expressions that are well known to an American speaker. Given that this is just a small sample of all the phrases that are made us of in the transmission, Jackendoff (1995) estimates that every speaker must have thousands of such word combinations stored in

their mind and that their number could thus more or less equate to that of single words. From such an observation stems the proposal for a new model of the lexicon that encompasses also these recurrent phrasal expressions, which we will explain in more detail below.

The received wisdom on linguistic creativity, and above all on single words being the units of this process, must then be revised by accounting for the interplay between formulaic, pre-constructed expressions on the one hand and phrases built on the fly on the other. Just like alternation between automatic and *ex novo* generated processes is observed in other types of behavior, including gestural, vocal and motor (Koestler 1967; Van Lancker & Cummings 1999), creativity and fixedness emerge as the two complementary roots of discourse creation (Bolinger 1976; Tannen 1989: 3; Sinclair 1991). On this subject, Hopper (1988) talks about *a priori* and *emergent* aspects of grammar, respectively. Lounsbury (1963: 561) describes *ad hoc* constructions and other combinations that are *"familiar and employed as a whole unit"* as different behavioral events that have a different psychological status in linguistic production. Noteworthy, the observation of highly recurrent chunks in everyday language dates back to the middle of XIX century, when Hughlings Jackson finds out that aphasic patients are unable to construct novel sentences, but still capable of remembering rhymes, routine greetings, prayers and so on. Saussure himself (1916/1966) describes the formation of complex expressions accessed as wholes when they are formed by common and frequent words:

> "when a compound concept is expressed by a succession of very common significant units, the mind gives up analysis – it takes a short cut – and applies the concept to the whole cluster of signs, which then becomes a simple unit" (Saussure 1916/1966: 177).

Jespersen (1924/1976) observes that language would be difficult to learn and to handle if its speakers had to remember every single item separately. Similarly, Bolinger (1976) regards as more convincing the idea of complex units stored in the speaker's mind that are then assembled via rules during sentence generation. His stance derives from reflection on the wide memory storage the human brain is effectively capable of. To say it with his own words, *"speakers do at least as much remembering as they do putting together"* (Bolinger 1976: 2).

A fundamental contribution to this issue comes from Sinclair (1991: 109 ff.), with his juxtaposition of an *open-choice principle* and an *idiom principle*. The first one conceives

language in a "slot-and-filler" perspective: texts are composed of slots that have to be filled from a lexicon. At each slot, any word is virtually possible, with the unique constraint of grammaticalness. Please note that this way of seeing texts is typical of most traditional and generative grammars. In a sentence tree representation like the one depicted above (Figure 1), each node corresponds to a choice point. Conversely, the idiom principle states that

> "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments". (Sinclair 1991: 110)

The idiom principle appears to be motivated by a series of extralinguistic and linguistic considerations. It may reflect the tendency of typical situations to recur in human affairs or a more general principle of economy or least effort. After all, Sinclair observes, the structure of extralinguistic reality inevitably influences the organization of language: for example, things that physically occur together or concepts belonging to the same field are more likely to be mentioned together. From a more linguistically-grounded angle, the choice of a certain register significantly constraints the set of words that can be used in a sentence. We must therefore assume that linguistic production and comprehension in their entirety cannot be satisfactorily accounted for with an open-choice principle. Let us consider the lexical bundle *of course*, which apparently consists in the simultaneous choice of two words, but actually operates as a unique complex word that always occurs in the same form, without any kind of variation. Furthermore, the behavior of its constituents is different from what we observe when they occur in other contexts and without each other. *Of* usually occurs after the noun head of a nominal group or in quantifiers like *plenty of* and in an open-choice model can be followed by any nominal group. At the same time, *course* does not behave like the countable noun listed in dictionary entries, since it would require a determiner in order to be grammatical. The point Sinclair (1991: 111) wants to make is that the meaning and the function of each component is not, or not simply, a property of the singular word itself, but of the whole phrase. His proposal is to store expressions like *of course*, but also other instances in which a single lexical choice involves more than one word, like idioms, proverbs, clichés and phrasal verbs, in the lexicon together with compounds. Indeed in all these expressions the constituent elements seem to have lost their semantic identity. In a similar way to Erman and Warren (2000), Sinclair goes on to observe a variety of features that characterize pre-constructed linguistic

chunks. First of all, many phrases can be extended and modified in an indeterminate way. An expression like *set eyes on* seems to frequently co-occur with a pronoun subject or temporal adverbials including *never, the moment* or *the first time*. It's worth asking whether such elements indeed represent a part of the phrase or are simply instances of collocational attraction (see below). Other phrases permit internal lexical variation, like *in some cases/in some instances*, together with variation of the reciprocal order of their parts: *set* X *on fire*/ *set fire to* X. Syntactic variation is possible too: in the phrase *it's not in his nature to…*, while *it, nature* and the prepositions emerge as the fixed elements, the verb can be inflected, the negation could be replaced by something like *hardly* or *scarcely*, and *his* could be substituted by any kind of possessive pronoun or genitive noun. Moreover, some words seem to co-occur more frequently than chance and thus to be strongly attracted, as in *dear friend, hard work, hard luck, good luck* and so forth. These distributional preferences emerge also at the syntactic level: a verb like *set about*, in the sense of "inaugurate", is normally followed by verbs in the *–ing* form, which are also usually transitive. Wray (1992) advocates a dual-processing model, in which the choice between analytical and holistic processing depends on the demands of the linguistic material and the communicative situation. Consequently, holistic retrieval is not confined to those expressions that cannot be built by means of rules (e.g. idioms), but can also be applied to combinations that are perfectly built with an analytical approach.

As Katz (1973) and Nunberg et al. (1994) have noted, attempts to thoroughly analyze and characterize the structure of formulaic expressions and to assign them a place in current linguistic models have met with mixed success. Prejudiced views have regarded them as *"inferior speech"* (Hughlings Jackson 1874; Redfern 1989) or a *"lazy solution to linguistic selection"* (Drew & Holt 1988). For this reason Fillmore et al. (1988: 534) make a plea for *"serious grammatical consideration of the 'realm of idiomaticity in a language'"*. Notwithstanding the abundant terminology that has been proposed for these phenomena, we will refer to such pervasive prefabricated combinations as *multiword expressions* (*MWEs*).

## 1.2. Multiword Expressions: Definition and properties

After giving a general definition of the heterogeneous class of MWEs, we briefly present the main classifications that have been proposed in the linguistic research of XX

and XXI centuries (cf. Bartsch 2004; Masini 2012). This section is followed by an extensive digression on *idioms*, which constitute the subject matter of the present work.

Multiword expressions, to adopt a terminology proposed by Zgusta (1967) and reemployed in recent studies (Sag et al. 2001; Calzolari et al. 2002; Gries 2008), can be described as *sequences of words acting as single units at some level of linguistic analysis* (Calzolari et al. 2002). The linguistic phenomena usually comprised under this label include:

- **collocations**, namely sequences of habitually co-occurring words that are at least partially compositional (Frege 1892), in that *"each lexical constituent is also a semantic constituent"* (Cruse 1986: 40; cf. Masini 2012). Some examples are *strong tea*, *torrential rain* or *heavy smoker*;[1]
- **idiomatic expressions** or **idioms**, both fully lexicalized (*get the sack*, *pull strings*) and with free slots (*grit* X*'s teeth*, *jog* X*'s memory*)*;* they are different from collocations in the strict sense, since they are noncompositional and characterized by high lexicosyntactic rigidity, figurativity and proverbiality (Katz & Postal 1963; Wasow et al. 1984; Cacciari & Glucksberg 1991; Nunberg et al. 1994; Jackendoff 1995);
- **light verb constructions**, formed from a commonly used verb and a direct object NP, such as *make a decision*, *have a look* or *give a groan*. In these

[1] Actually, there is a great terminological chaos about *collocations*, since the word is sometimes used as an all-embracing definition for every constrained combination of words (cf. Manning & Schütze 1999: 139 ff.) and sometimes with a narrower scope (Masini 2009). Firth (1957b: 194) employs it in his contextual theory of meaning, explicable with his well known motto *"You shall know a word by the company it keeps"* (Firth 1957a: 11). In this sense, he claims, one of the fundamental ways to assess the meaning of a given word is to look at its *collocational behavior*, defined in terms of the syntagmatic contexts in which it most frequently occurs. Sinclair (1991: 170) claims that collocation in a wider sense means *"the occurrence of two or more words within a short space of each other in a text"*, while in a more restricted sense it denotes a *"frequently repeated"* combination of lexemes. Evert (2008) advocates a clear distinction between *empirical collocations*, which are just statistically relevant word combinations in a corpus (e.g. word pairs that co-occur more often than chance), like *to write a book* or *dear friend*, and effective *lexical collocations* belonging to MWEs, which assume a phraseological status. In lexical collocations, one of the components may be used only in combination with the other (*acqua potabile* "drinking water", where *potabile* only co-occurs with *acqua*) or there could be available synonymous expressions, which nevertheless are never used (*sito ufficiale* "official site" vs. **sito autorizzato*) (Tiberii 2012; Squillante 2014). Masini (2009) tells apart collocations and *preferential combinations*. In the first case (*aprire un conto* "open a bank account") the presence of a word (*conto*) mandatorily requires the other (*aprire*). In the second one, (*pioggia torrenziale* "torrential rain") one of the words (*pioggia*) preferentially combines with the other (*torrenziale*), but equivalent combinations are in any case available (*pioggia intensa* "heavy rain" and the like).

combinations, it is mainly the noun component to contribute to the meaning of the whole string, while the verb semantics is somehow bleached (Kearns 2002; Butt 2003). This is confirmed by the fact that we could replace a LVC (*make a decision*) with a single verb that is morphologically related to the noun constituent (*decide*) while keeping the whole meaning intact (Fazly & Stevenson 2008);

- **irreversible binomials** (*black and white*, *fish and chips*);
- **proverbs** (*the early bird catches the worm*, *better late than never*);
- **speech formulae** (*what's up*, *thank you*);
- **exclamations** (*what the hell*);
- **quotes** and titles from poetry, music, books and so forth

As regards the criteria that have been advanced in the literature for defining a given combination as a multiword unit, they are both structural and semantic in nature (Sag et al. 2001; Calzolari et al. 2002; Wray 2002; Fazly & Stevenson 2008; Masini 2012). Structurally speaking, the fundamental prerequisite is *lexicosyntactic fixedness*, which means that:

(a) the component words cannot be replaced by equivalent ones;
(b) the single constituents are not modifiable and separable;
(c) the components are disposed in a fixed order;
(d) the disposition of the words can sometimes violate general syntactic patterns or rules;

It must be underlined that the coexistence of all this criteria is not mandatory, maybe save (a). Moreover, every criterion does not represent a clear-cut choice, but manifests itself in a gradient manner. If we take a literal combination like *give a present*, its meaning would practically remain the same if we replaced *present* with a more or less semantically related word like *gift.* On the contrary, changing idioms like *shoot the breeze* and *kick the bucket* into *shoot the wind* and *boot the bucket* does not preserve their figurative interpretation anymore. Nevertheless, this is also true for some literal combinations, namely collocations, that are literal, but anyway subject to restricted lexical selection: *authorized site* would sound quite strange and atypical for an English native speaker, while

*official site*, though semantically equivalent and literal in the same way, would result perfectly acceptable. Moving to (b) and (c), as we will note over and over throughout the present work, we should note that the constituents of a MWE are often separable, e.g. via adjectival or adverbial insertion or via various syntactic movements, though the degree of variability of a given multiword unit with respect to these phenomena is not predictable. **My grandpa kicked the atrocious bucket* or **The bucket that my grandpa kicked was atrocious* would be ungrammatical, while saying something like *The wounded man gave a terrible groan* or *The groan the wounded man gave was terrible* would not be so puzzling. The classes that maybe respect (c) most rigidly are irreversible binomials, e.g. *black and white* vs. **white and black,* but also some idioms, e.g. *spic and span* vs. **span and spic.* Violation of general syntactic patterns as predicted by (d) can be traced in multiwords like *in the know* or *down and dirty*.

From a semantic perspective, the (more or less) discriminating criteria are:

(a) reduced or lack of compositionality;
(b) functional equivalence with simplex words;
(c) high degree of conventionality;

According to Frege's (1892) *principle of compositionality,* the meaning of a complex expression is a function of the meaning of its constituent parts and of the relations among them. This perfectly applies to *literal combinations* like *white car*, which is something white and is a car at the same time, or to a literal sentence like *The old man wrote a letter.* In this case, the meaning of the noun phrase is obtained by combining *the*, *old* and *man* according to the grammar rules that govern the generation of a noun phrase, the meaning of the verb phrase is obtained by combining *wrote*, *a* and *letter* once more according to the grammar rules for verb phrases and the meaning of the whole sentence is obtained by combining the noun phrase meaning and the verb phrases meaning. In the case of a *light verb construction* like *give a try*, while *try* receives a literal interpretation, *give* is not interpreted so and even seems semantically dummy for the interpretation of the whole meaning. This is also corroborated by the observation that we could replace the entire expression with the simple verb *to try.* Let's now consider an *idiom* like *take the bull by the horns*. In this case, we cannot arrive at the figurative meaning "to face a difficult situation in a brave and direct way" by simply combining the meaning of its component words according to the rules of English grammar. In this specific case, the literal meaning

represents a sort of archetypical and symbolic instance of the figurative meaning, but this connection is obviously not automatically accessible for a native speaker unless he/she has learnt it. In any case, the idiom we just mentioned possesses both a literal interpretation, unrelated or figuratively related to its idiomatic meaning, and an idiomatic one. Hence, we define it an *ambiguous idiom*. For other idioms, namely *non-ambiguous* ones, noncompositionality can even result in a total absence of literality. We define *literality* as the extent to which an idiom displays a plausible literal interpretation (Popiel & McRae 1998). To provide some examples, *go bananas* and *shoot the breeze* cannot be read literally. In the first case, this is due to syntactic anomaly, since *go* is an intransitive verb that cannot normally co-occur with a direct complement; as for *shoot the breeze*, the anomaly is semantic in nature, insofar as selectional restrictions for both *shoot* and *breeze* are violated: breezes are not something that can be shot. All in all, native speakers have to learn the meaning of a given idiomatic combination to properly use it and understand it in a context. Finally, *give attention* appears as an in-between case, in that *attention* preserves its meaning, while the verb *give* does not denote a physical transfer, as it prototipically does. What we witness here is a metaphorical extension of this prototypical and concrete meaning, that becomes a transfer of a psychological and cognitive state. As a consequence, when speakers process *abstract combinations* like this one, they have to reverse-engineer the metaphor involved and trace back the analogy between the basic and the extended meaning of the component words (Lakoff & Johnson 1980; Newman 1996).

As regards point (c), we observe that a certain degree of *institutionalization* is involved in any MWE. We define it as the process whereby *"a combination of words becomes accepted as a conventional semantic unit"* (Fazly & Stevenson 2008). According to usage-based models (cf. Goldberg 2006; Bybee 2010; see Chapter 2) we expect a positive correlation between conventionalization and semantic idiosyncrasy: the more institutionalized an expression, the more entrenched it will become in the speaker's mind and the more likely the speaker will access it as a whole. Crucially, from a corpus-based perspective, institutionalization cannot be approximated by mere frequency, since a combination can occur with high frequency just by chance and because its components are highly frequent by themselves. We must thus resort to collocational measures like *Pointwise Mutual Information* (*PMI*; Church et al. 1991) that calculate the degree of statistical association between two words by comparing their expected and observed frequencies.

We have just observed that different kinds of MWEs display different degrees of

compositionality. We could therefore arrange them on a *semantic idiosyncrasy continuum*, that spans from literal combinations (*white car, write a letter*) to abstract combinations (*give attention, put a price*), light verb constructions (*give a try, give a groan*) and, last but not least, idioms (*give a whirl, take the bull by the horns*):

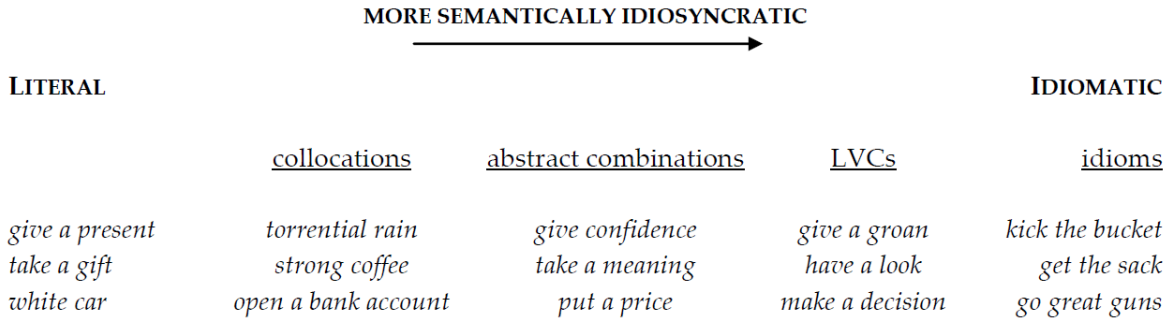



**Figure 2:** MWEs projected on the semantic idiosyncrasy continuum (adapted from Fazly & Stevenson 2008)

As we will see, the phenomenology of MWEs has also been addressed in psycholinguistic and computational terms. From the first point of view, the core notion is the *holistic retrieval* from the mental lexicon in both production and comprehension (Sinclair 1991; Wray 2002):

> "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, *stored and retrieved whole from memory at the time of use*, rather than being subject to generation or analysis by the language grammar" (Wray 2002: 9, italics mine)

From a computational and probabilistic viewpoint, we can define MWEs as combination of words that co-occur more often than chance (Manning & Schütze 1999).

Anyway, as Wray (2002: 8 ff.) reports, multiwords and pre-constructed sequences have been labeled with a plethora of denominations according to the different fields of study. Following her description, we list here below all (or almost all) the existing nomenclature on fixed expressions:

| |
|---|
| amalgams – automatic – chunks – clichés – co-ordinate constructions – collocations – complex lexemes – composites – conventionalized forms – Fixed Expressions including Idioms – fixed expressions – formulaic language – formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen phrases – gambits – gestalt – holistic – holophrases – idiomatic – idioms – irregular – lexical simplex – lexical(ized) |

phrases – lexicalized sentence stems – listemes – multiword items/units – multiword lexical phenomena – noncompositional – noncomputational – nonproductive – nonpropositional – petrifications – phrasemes – praxons – preassembled speech –precoded conventionalized routines – prefabricated routines and patterns – ready-made expressions – ready-made utterances – recurring utterances – rote – routine formulae – schemata – semipreconstructed phrases that constitute single choices – sentence builders – set phrases – stable and familiar expressions with specialized subsenses – stereotyped phrases – stereotypes – stock utterances – synthetic – unanalyzed chunks of speech – unanalyzed multiword chunks – units

**Figure 3:** terms used to describe aspects of formulaicity (Wray 2002: 9)

While some terms refer to the same elements, there are also cases in which the same definition is used by different scholars with different meanings. In general, the wide and varied class of multiword units has been carved and divided in many ways, *"none of which"* nonetheless *"seems fully to capture the essence of the wider whole"* (Wray 2002: 8).

## 1.3. Proposed classifications of MWEs

In pre-structuralist era, Paul (1880/1920) and Brèal (1897/1904) reflect on the existence of fixed locutions, which appear to be characterized by non substitutable component elements and lack of semantic decomposability. Sweet calls *group-compounds* those constructions like *son-in-law* that seem to exist on an intermediate level between compounds and free combinations (Graffi 1991: 248). In the Saussurean *Cours* (1916/1922), one of the main questions that come to the fore is the combinatorial freedom of phrases and their more or less exclusive belonging to the *parole.* On the one hand, the sentence is the syntagmatic combination *par excellence* and it belongs to the *parole*. On the other hand, the basic characteristic of the *parole* is the total combinatorial freedom of its elements and this does not seem to apply to every phrase: there are plenty of "*locutions toutes faites"* in language (Saussure 1916/1922: 172), like *à quoi bon?* "what for?" or *forcer la main à quelqu'un* "force someone's hand" that do not exhibit free choice of components and are not consequently attributable to the *parole*. Another fundamental contribution within Genevan structuralism to the study of multiword expressions comes from Bally (1909/1951: 66-87), who paves the way for all the following studies in phraseology. In a continuum spanning from freely combinable words to indecomposable units, we find *"locutions phraséologiques"* in the middle. They can be divided as:

- *séries phraséologiques*, with relative internal cohesion, in that internal elements preserve a certain degree of autonomy (*gravement malade* "seriously ill", *prendre une décision* "make a decision")
- *unités phraséologiques*, with total internal cohesion, that exhibit *a)* fixedness in the order of the constituents; *b)* inseparable and irreplaceable components; *c)* archaisms or words that never occur by themselves (*en guise de* "in the manner of"); *d)* equivalence in meaning with simplex words (*manière d'agir* vs. *procédé* "behavior"); *e)* loss of meaning of the component words;

After the works of Bally, phraseology meets with success in Russian linguistics, in particular with Mel'čuk (1998). He uses *phraseme* as overarching definition for all MWEs, which are further classified as:

- *pragmatemes*, that include formulae and common sayings;
- *idioms*
- *collocations*
- *quasi-idioms*, that encompass the meaning of the single lexemes with an additional unpredictable element (*to start a family*, *shopping centre*)

Cruse (1986) distinguishes complex lexical units depending on their *semantic opacity*. The most opaque class are *idioms*, defined as lexical complexes that are semantically simplex (Cruse 1986: 37). Irreversible binomials (*fish and chips, bacon and eggs*) are instances of semi-opaque elements, since their subparts contribute to their overall meaning, while *collocations (fine weather, torrential rain*), as we just said in the paragraph above, are depicted as "*sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent"* (Cruse 1986: 40). In his list of criteria for recognizing a MWE, Zgusta (1971: 151) adds that *"a multiword lexical unit must perform, in a sentence (syntagmatically) and in the lexicon [...] (paradigmatically) the same syntactic and onomasiological function as a morphologically more simple unit which frequently coincides with the word, e.g. in the Indo-European languages"*. This automatically excludes proverbs, clichés and quotes, which he classifies separately as *set groups of words*.

A significant contribution to the study of MWE within a corpus-linguistic perspective comes from Moon (1998), who exploits the *Oxford Hector Pilot Corpus* (18 million of words) to analyze a wide range of what she calls *FEI* (*Fixed Expressions and Idioms*), comprising collocations, proverbs, formulae and idioms. Just like Bolinger (1977: 168) observes that the boundaries between an idiom and a collocation and between a collocation and a freely constructed phrase are blurred, Moon (1998) puts FEI in a continuum spanning from compositional to non-compositional groups of words which is more complex than a binary compositional/non-compositional distinction. To isolate a FEI she proposes three main criteria, namely *institutionalization, lexical and grammatical rigidity* and *non-compositionality* and other secondary criteria, such as the division of each expression in at least two orthographical words or the coincidence of a FEI with a syntactic or grammatical unit (a phrase or a sentence). The classification she proposes is based on the reason why a given word-combination should be lexicographically considered a holistic unit, i.e. whether the sequence is anomalous from the viewpoint of lexicogrammar, pragmatics or semantics (Moon 1998: 18-19). In the first case we have *anomalous collocations*, in turn divided into:

- *ill-formed collocations*, that are extragrammatical (cf. Fillmore et al. 1988: 505), like *by and large, of course* and *thank you*;
- *cranberry collocations*, that contain elements which do not appear elsewhere, like *on behalf of* and *to and fro*;
- *defective collocations*, which are not fully compositional, such as *beg the question* or *in effect*;
- *phraseological collocations*, that more or less correspond to Fillmore et al.'s (1988) *formal idioms*, i.e. strings that have alternative versions but are not fully productive, like *in action/into action/out of action* or *to a ... degree/to a ... extent*;

Pragmatically anomalous strings are *formulae* (*alive and well*), *sayings* (*an eye of an eye*), *proverbs* and *institutionalized similes* (*as old as the hills*).

If the anomalies are located in semantics, Moon (1998: 19 ff.) talks of *metaphors*:

- *transparent metaphors* are institutionalized but allow the speaker/hearer to reverse-engineer them by relying on world knowledge, like *alarm bells ring* or *behind someone's back*;
- *semi-transparent metaphors* require some additional knowledge for their interpretation, such as *throw in the towel* or *grasp the nettle*;
- *opaque metaphors* are nothing but pure idioms, e.g. *kick the bucket*, *red herring*;

In a similar fashion to a series of studies coming from discourse analysis, Moon (1998: 217) dwells also on the discourse functions of FEI. While, for instance, some expressions are informational in nature, in that they state propositions or convey information (e.g. *rub shoulders with*, *catch sight of something*), some other are evaluative, insofar as they convey the speaker's evaluation or attitude towards the denoted reality (e.g. *kid's stuff*, *another kettle of fish*) and so on. Previously, Nattinger and DeCarrico (1992) already propose a functional distinction between *social interactions*, *necessary topics* and *discourse devices*, while Aijmer (1996) describes a taxonomy for conversational routines. Anyway, these classifications in terms of discourse-functional criteria tackle just a part of the role that MWEs have in language. As Cowie (1988: 134) affirms, there are fixed expressions that just contribute to the referential or propositional meaning of a sentence (what he calls *composites*) without any salient discourse function.

Within the literature on *Natural Language Processing* (*NLP*), Becker (1975), unsatisfied with the too abstract and idealized treatment of language that is proper to previous linguistic theories (e.g. generativism), proposes to switch the focus on the *phrasal lexicon* we actually draw on in real linguistic practice. In his view, linguistic processing is compositional in the sense that, once the speaker has selected the message to convey, he/she extracts from the phrasal lexicon the patterns that best express it and stitch them in a sentence. Crucially, this patterns are often fixed sequences more complex than single words. The next step is a so-called *generative gap-filling* (Becker 1975: 62) in which the speaker adapts and complements this pattern-structured sentence with the details of the situation at hand. This lexicon is characterized by a complex taxonomy. We have:

- *polywords*, that admit no variability and are interchangeable with single words, like *the oldest profession* or *for good*;

- *phrasal constraints*, composed of a small number of words, some of which constrain the variability of the others, as in *by pure coincidence*/*by sheer coincidence*;
- *deictic locutions*, that present discourse function, such as *for that matter* or *that's all*;
- *sentence builders*, that can be complex up to sentence length and provide the schema for expressing an idea, like *give someone a song and dance about something* with the meaning of "giving a long explanation about something"
- *situational utterances*, that is fixed sentences like *How can I ever repay you?*
- *verbatim texts* like proverbs and quotes

Sag et al. (2001) underline the urgency to develop sound NLP systems that take all the properties of MWEs into account. One of the main reasons to do so is that MWEs are pervasive in any language and textual genre and new exemplars come into existence on a daily basis, so lexical resources in general and computational lexica must be kept up to date to soundly perform NLP tasks (cf. Fazly et al. 2009). To provide some examples, collocational restrictions could be overlooked by a natural language generation algorithm that relies on general compositional methods, with a consequent overgeneration problem: **telephone cabinet* or **telephone closet* would be produced beside correct collocations like *telephone booth* (American English) and *telephone box* (British/Australian English). While processing an idiomatic string like *red herring*, it must be highlighted that its meaning is not at all related with that of *red* or *herring*. Finally, MWEs that are formed by violation of conventional grammatical rules, such as *down and dirty* or *in the know*, can represent a problem for parsing algorithms. Many NLP models approach multiwords in terms of *words-with-spaces*. While this can successfully deal with structurally rigid combinations like *by and large*, Sag et al. (2001) note that it is not sufficient for flexible multiwords like verb-particle constructions, that undergo different interpretations according to the order in which the constituents appear (cf. *look up the tower* meaning both "glance up at the tower" and "consult a reference book about the tower" with *look the tower up* which is only figurative) or certain idiom classes (see below) that allow morphosyntactic variability, like *hold fire* that can manifest as *held fire* or *hold one's fire* et cetera. As an additional shortcoming, they observe an inability of such systems to abstract over the common characteristics of a given set of MWE, like *take a walk*, *take a hike*, *take*

*a trip* etc., which would lead to greater generality and predictive capacity. Instead of doing so, these algorithms treat every case as an isolated entry, with increased preprocessing costs. Drawing on Bauer (1983), they divide MWEs into *lexicalized phrases*, which are idiosyncratic in their syntax or semantics or contain words that never occur by themselves, and *institutionalized phrases*, that are compositional but occur with considerably high frequency. Lexicalized phrases include:

- *fixed expressions*, that are fully lexicalized and never undergo morphosyntactic variation or internal modification, like *by and large*, *in short*, *ad hoc*, *ad nauseum* and names like *Palo Alto, Alta Vista* etc.
- *semi-fixed expressions*, that allow some variation, e.g. in inflection or determiner selection, like non-decomposable idioms (see below; Nunberg et al. 1994), compound nominals (*car park, part of speech*) and proper names (*the San Francisco 49ers*, with *San Francisco* being erasable)
- *syntactically-flexible expressions*, including verb-particle constructions (*call someone up*/*call up someone*), decomposable idioms (see below), and light-verb constructions (*make a mistake, give a demo*)

*Institutionalized phrases*, on the other hand, are compositional in their semantics and syntax, but statistically idiosyncratic and hence undergo conventionalization. *Traffic light*, for example, is perfectly compositional but the same concept cannot be expressed by synonymous expressions like *traffic director* or *intersection regulator*. These alternative versions that are never or almost never encountered are labeled as *anti-collocations* (Pearce 2001). Please note that Sag and colleagues (2001) use the term *collocation* to refer to any sequence that is statistically significant. In other words, collocations are not only any kind of MWEs but also all predictably frequent compositional phrases, like *sell-house*.

## 1.4. Narrowing the focus on idioms

### *1.4.1. Definition and properties*

After dealing with the classifications and descriptions that have been proposed on MWEs in general, we now restrict our focus on idioms. First of all, the term *idiom* can refer to:

1. *"A language, especially a person or people's own language; the distinctive form of speech of a particular people or country"* and in a narrower sense *"a form of language limited to or distinctive of a particular area, category of people, period of time or context"* (OED, s.v. *idiom*, senses 2a and 2b)
2. *"A form of expression, grammatical construction, phrase, etc. used in a distinctive way in a particular language, dialect or language variety;* spec. *a group of words established by usage as having a meaning not deducible from the meaning of the individual words"* (OED, s.v. *idiom*, sense 3)

The relevant sense for the study at hand is the second one (Cacciari & Tabossi 1993: xi; Wulff 2008: 9), although the two meanings are to a certain extent connected (Fillmore et al. 1988). As native speakers of our language, we can have an intuitive perception of what idioms are, especially in light of their pervasiveness in everyday communication: Searle (1975: 50) states that speakers seem to follow the implicit rule *"Speak idiomatically unless there is some special reason not to"*. Typical English idioms could be *kick the bucket* "to die", *spill the beans* "to divulge a secret" or *shoot the breeze* "to chat idly", while for Italian we could cite *tirare le cuoia* "to die" (literally untranslatable and roughly corresponding to *kick the bucket*), *alzare il gomito* "to drink too much (alcohol)" (lit. "to raise the elbow") or *mettere le carte in tavola* "to lay one's cards on the table, to show one's intentions" (lit. "to put the cards on the table"). Nevertheless, as we go deeper in theoretical reflection and categorization attempts, we find out that idiomaticity actually constitutes a *multifactorial* concept (Wulff 2008). That is to say, exhaustive criteria that neatly tell apart idiomatic expressions and all other kinds of figurative language on the one hand and idioms and other MWEs on the other bring into play several semantic, syntactic and psycholinguistic considerations (Cacciari & Glucksberg 1991; Nunberg et al. 1994; Wulff 2008).

Nunberg et al. (1994: 492) state that idioms indeed "*occupy a region in a multidimensional lexical space"* and highlight a series of orthogonal properties, not all of them being indispensable, that nonetheless seem to be distinctive for this heterogeneous class of expressions, namely:

1. **Inflexibility –** in contrast with freely combining expressions, every idiom allows just a restricted set of morphosyntactic operations and, most importantly, in an

unpredictable way. For an English native speaker, *shoot the breeze* and *kick the bucket* would lose their idiomatic sense if they underwent adjective or attributive insertion (a), determiner change (b), passivization (c) or relativization (d), while all these morphosyntactic variations would result more acceptable for *spill the beans*;

2. **Conventionality** – i.e. their meaning can't be predicted just knowing the conventions that govern the use of their components when they appear by themselves and outside the expression. In other words, what immediately catches the eye about an idiomatic expression is that its meaning is not a function of the meaning of its constituent parts (Fraser 1970; Katz 1973; Bobrow & Bell 1973; Wasow et al. 1983; Cacciari & Glucksberg 1991). In analyzing *shoot the breeze*, even if we know the meaning of *shoot*, *the* and *breeze*, we cannot arrive at the idiomatic sense of "chatting idly" just combining the meaning of the three words in a normal compositional way, but we have to learn it as it is, for example by experiencing the given expression in a context. It's worth noting, on the other hand, that talking about compositionality is not so straightforward. First of all, the component parts, the rules and the functions at stake in building a complex expression are not always clearly discernible (Casadei 1996). Secondly, it is practically impossible to define what the absolute meaning of the single words is independently from a given context, in accordance with Firth's (1957b) maxim (see also Hanks 2013). Finally, "non-compositional" appears a too generic label for certain idioms: if there doesn't seem to be any kind of mapping, neither literal nor metaphorical, between the words in *kick the bucket* and the idiomatic sense "to die", in cases like *spill the beans*, we can interpret *spill* as metaphorically referring to the act of divulging and *beans* as standing for the secrets. Some idioms are therefore not compositional in the traditional way, but nonetheless *semantically decomposable*, with a more or less metaphorical mapping between the constituents and their idiomatic referents (Nunberg et al. 1994). In any case, this *semantically idiosyncratic* nature of idioms calls for more reasonable models of language production and comprehension than the traditional "compositionality-oriented" ones (Bobrow & Bell 1973; Cacciari & Tabossi 1988; Gibbs, Nayak, & Cutting 1989; Swinney & Cutler 1979; Titone & Coninne 1999) and, in building lexicographical resources and computational lexica, it requires a specific entry for each idiom, that associates it with its unpredictable meaning (Fazly et al. 2009).

(1) (a) **John and Mary shot the pleasant breeze*
**Our neighbour kicked the sad bucket last year*
*Jim spilled the theft beans while interrogated by the police*

(b) **John and Mary shot a breeze*
**Our neighbour kicked a bucket last year*
*Jim spilled some beans while interrogated by the police*

(c) **The breeze was shot by John and Mary*
**The bucket was kicked by our neighbour last year*
*The beans were spilled by Jim while he was interrogated by the police*

(d) **The breeze that John and Mary shot was pleasant*
**The bucket that our neighbour kicked last year was painful*
*The beans that Jim spilled while interrogated caused him trouble*

As for (a), it must be precised that an adjective modification like *kick the proverbial bucket* would still be acceptable, since it would constitute a metalinguistic comment about the idiom itself and not an actual modification of an internal part of the expression (Cacciari & Glucksberg 1991). The morphosyntactic form in which an idiom preferentially occurs, with respect to gender, number and definiteness of the component noun(s), diathesis of the component verb(s) and the like, is called *canonical form* (Glucksberg 1993; Riehemann 2001; Grant 2005; Fazly et al. 2009). Lexical flexibility represents another discriminating factor. Replacing *spill the beans* with *spread the beans* or *spill the peas* results in the loss of the idiomatic meaning (Gibbs, Nayak & Cutting 1989; for a computational treatment of this aspect cf. Lin 1999 and Fazly et al. 2009). Corpus studies have in fact demonstrated that some kind of lexical variation is permitted: Moon (1998) finds *kick the pail* and *kick the can* as lexical variants of *kick the bucket*. Anyway such restrictions are not predictable and semantically constrained with respect to the situation denoted. As Glucksberg (1993: 7) notes, since *break the ice* refers to an event that chills out an awkward social situation, this metaphorical ice is not something that could be *crushed*, *grinded* or *shaved* for an English native speaker, while *crack the ice* or *melt the ice* would sound as acceptable substitutions. This kind of lexical restrictions have important connections with the way idioms are processed (see below). Idioms in a running text are in fact processed linguistically, even when this processing is not necessary for determining their figurative

meaning. The literal meaning of the component words preserves thus an important role in idioms production and comprehension.

3. **Figuration –** be it a metaphor (*take the bull by the horns*), a metonymy (*lend a hand*) or a hyperbole (*not worth the paper it's printed on*), every idiom involves figuration in some extent. Even when speakers are not able to go back to the origin of such figurative readings and explain why a given meaning is expressed by a given metaphor, they still perceive that some kind of metaphorical or metonymical extension is brought into play. As psycholinguistic evidence confirms, no or just a few speaker would regard, for instance, the word *bullet* in the idiom *bite a bullet* as a mere homonym of the word meaning "projectile": the presence of figuration is therefore almost always felt (Gibbs 1990), contrary to Kiparsky's (1976: 79) claim that many idioms are no longer perceived as metaphorical. It is also true, anyway, that figuration does not apply to certain idioms that contain words never occurring in isolation, like *by dint of,* for which we could not distinguish between a literal and a figurative intepretation of *dint.*
4. **Proverbiality –** idioms usually refer to common and recurrent situations by means of scenes involving familiar things and actions that indirectly represent them. The acts of talking in an informal way or revealing something secret, for instance, are represented by the acts of *chewing fat* or *spilling beans,* that constitute objects and actions quite typical and familiar in everyday life. It goes without saying that this represents just a tendency, since we can have idioms like *be in seventh heaven*, *there's method to someone's madness* or *at sixes and sevens* which don't refer to homey and concrete things at all.
5. **Informality –** they are usually associated with informal registers. Nonetheless, idioms with a literary flavor also exist, like *render unto Caesar.*
6. **Affect –** they often imply an affective evaluation of the reality they refer to; for this reason, actions that are regarded as emotionally neutral within a certain culture, like taking a bus or buying food, are not denoted by idioms. This affective value of idioms is at the root of their recent exploitation in sentiment analysis algorithms (cf. Williams et al. 2015).

To differentiate idiomatic expressions from other instances of figurative language (e.g. metaphors, proverbs and clichés), we must notice that while idioms have a unique meaning that can adapt to various contexts without being changed, both frozen and less

conventionalized metaphors can convey different meanings when used in different situations. *John is an elephant* could in effect mean either that he is big or that he is clumsy and bumbling according to the context of use. Moreover, we can generate a new metaphor at whatever time, while idioms are fixed, conventionalized and stored in our lexicon and we could not creatively combine words to produce a novel one on the fly (Konopka & Bock 2009; Cacciari & Papagno 2012). While producing and comprehending metaphors requires a categorization process (Glucksberg & Keysar 1990; Cacciari & Glucksberg 1994; see below), idioms are accessed and retrieved from semantic memory (Cacciari & Papagno 2012; Cacciari 2014). On the other hand, proverbs differ from idioms in that they are full sentences without temporal definition and are generally true both literally and figuratively (Turner & Katz 1997; Cacciari & Papagno 2012).

The conception of idiomaticity as a multidimensional phenomenon has gradually developed through the decades, the traditional treatment being mainly focused on the feature of non-compositionality (Katz & Postal 1963; Weinreich 1969; Chomsky 1980). Evidence from corpus-based (Wulff 2008) and psycholinguistic studies (Gibbs & Nayak 1989; Gibbs et al. 1989; McGlone et al 1994) has shown that speakers rely on a series of interrelated features to label a certain construction as idiomatic. Anyway, before seeing how idioms have been differently analyzed in a series of major studies, we must first linger on the effective notion and meaning of *idiomaticity*.

First of all, the term *idiomaticity* is related but not equivalent to *idiom* (Fernando 1996), in that it denotes the very property that encompasses *all* the expressions on the semantic idiosyncrasy continuum depicted above (Figure 2) and that is basically manifested in terms of non-compositionality and formal flexibility. Real idioms represent in this sense the prototypical instantiation of this property. To say it in Wray's (2002: 34) words, *"the feature ±idiom could be a defining variable in a typology of formulaic sequences along a continuum from fully bound to fully free".* In her work on idioms from a functional and discourse viewpoint, Fernando (1996: 38) employs the term to refer to any conventionalized multi-word item, be it non-literal or not. In her theory, habitual co-occurrence generates idiomatic expressions, but only those idiomatic expressions that acquire fixed order and elevated formal rigidity become real idioms. She then proposes a continuum of multiword units whose ends are occupied respectively by idioms and collocations and where compositionality does not emerge as a discriminating element, since both literal and non-literal expressions appear in both the idioms and the collocations class. Conversely, only variable elements appear in the collocations category. Similar is the

idiomaticity continuum described by Howarth (1998), that spans from *free combinations* to *restricted collocations*, *figurative idioms* and *pure idioms.* In this case, the non-idiomatic end is characterized by formal fixedness, while the idiomatic one is characterized by semantic non-compositionality. In the second place, we have started our analysis of idioms by saying that the two meanings of the term, namely (a) the language of a given area or people and (b) a typical, language-specific and non-compositional expression, appear after all related when defining the very concept of idiomaticity. In effect, Fillmore et al. (1988) draw on Makkai (1972) in separating *idioms of encoding* and *idioms of decoding*. While idioms of decoding cannot be deciphered with complete confidence unless the speaker has already learnt them, idioms of encoding are more or less comprehensible at first hearing, but, most importantly, an unaware speaker would not perceive them as conventional ways to express what they express. Traditional idioms are therefore of both the encoding and the decoding type, while collocations like *wide awake* or *answer the door* belong just to the encoding one, because they are perfectly compositional and understandable without prior experience, although not identifiable as institutionalized. Such a complex and dual conception of idiomaticity recalls Pawley and Syder's (1983: 193) quote on idiomaticity as that kind of nativelike selection that constraints all the virtually generable sentences in language:

> "only a small proportion of the total set of grammatical sentences are nativelike in form - in the sense of being readily acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be *'unidiomatic'*, 'odd' or 'foreignisms'" (italics mine)

### *1.4.2. Generative accounts of idioms*

Although isolated contributions in the '50s begin to address the transformational issues of idioms (Bar-Hillel 1955), the first major study on idioms within a generative perspective is conducted by Katz and Postal (1963). Starting from the by now ascertained equation between idiomaticity and non-compositionality, they move on to separating *lexical idioms* and *phrasal idioms.* Lexical idioms are non-compositional sequences of morphemes that are dominated by lowest level syntactic categories, like nouns, adjectives and verbs. Examples are compounds like *bari+tone* or *tele+phone,* for which we don't have

projection rules that take the readings of the single morphemes from the dictionary of the semantic component and amalgamate them, but the reading representing their sense is directly assigned to them. When such sequences are not dominated by lowest level syntactic categories we talk of *phrase idioms*, like *kick the bucket*. Their main question is how phrase idioms, the main focus of their study, can be handled within the semantic theory they adhere to (cf. Katz & Fodor 1963). First of all, this stance posits a set of syntactic rules and a lexicon constituting the syntactic component of a linguistic description. While lexical morphemes are listed in the lexicon, syntactic rules generate the phrase marker of a sentence, whose terminal elements are constituted by grammatically marked positions to be filled by lexical items. If the grammatical markings of a lexical item correspond to those of a terminal element in the phrase marker, the item is inserted in that position of the sentence. After that, *projection rules*, a central device in the semantic component of a language, assign the correct semantic interpretation to a lexical item in a sentence by selecting it among all the senses listed for that item in the lexicon. In such a framework, treating phrase idioms as lexical idioms, and so as representatives of lowest level syntactic categories (e.g. *kick the bucket* with the meaning "to die" would be regarded as a compound verb), poses problems for idiomatic strings that are also liable to literal interpretation, since the synactic and phonological component of linguistic description would be uneconomically complicated. Creating a specific entry "intransitive verb = *kick the bucket*" in the syntactic lexicon is in fact unnecessary, since the syntactic component is already supposed to generate the equivalent compositional string, in which each atomic constituent is a lexicon entry. An additional issue concerns the phonological component, which operates on the phrase marker of a sentence to give it a phonetic shape, including a stress pattern. Nonetheless, what we observe is that both an idiomatic and a literal version of *kick the bucket* have the same stress pattern. Consequently, they must correspond to equivalent syntactic descriptions and not to an intransitive verb on the one hand and to a verb plus noun phrase to the other. Katz and Postal (1963) therefore propose to divide the semantic dictionary into a lexical-item part and a phrase-idiom part. In the second one, the sequence of morphemes constituting an idiomatic stretch are associated with the constituent that dominates the idiomatic stretch in the phrase marker and that must receive the idiomatic meaning and with the semantic information itself. This bipartition in the dictionary of the semantic component is reflected in two different ways for lexical and phrase idioms to obtain semantic interpretation in a sentence. In a phrase marker *M*, a minimal element *e* consisting in a single morpheme or a lexical idiom obtains those

readings from its dictionary entry that are compatible with its syntactic structure in *M*. Otherwise, if we are dealing with a phrase idiom like *kick the bucket*, the semantic interpretation is directly assigned to the higher level constituent that dominates the string and not to the atomic components of the string. In this case, the entry for *kick the bucket* in the phrase-idiom dictionary section is associated with the consituent MV, which directly receives the interpretation "to die" in the phrase marker, resulting in the lexical entry "kick+the+bucket → MV → reading corresponding to the meaning 'to die'". Finally, an important reflection is carried out about the problem of transformational deficiency. Analyzing the following sentences:

(2) a. *Mary kicked the bucket.*
b. *The bucket was kicked by Mary.*

the fact that (b), differently from (a), can only be read literally is motivated by Katz and Postal (1963) In light of Klima's (unpublished, 1960-1961) theory of passive phrase markers. Klima affirms that a passive sentence is not derived from the same underlying phrase marker of an active sentence via passive transformation, but that it has a different phrase marker containing a Manner Adverbial component, terminally represented by a passive morpheme. The passive surface sentence is then obtained by substituting *by* plus the subject NP with the passive morpheme, placing the object NP in the subject position and adding some elements to the auxiliary constituent. Consequently, in the phrase marker of (a), the string *kick the bucket* is normally dominated by MV alone and it is therefore liable to both readings, while in the phrase marker of (b), MV dominates *kick the bucket* plus the passive morpheme: the condition required by the lexical entry "kick+the+bucket → MV → reading corresponding to the meaning 'to die'" is therefore not satisfied and the only available reading is the literal one.

Weinreich (1969: 42) frames his definition of idiom in Soviet phraseology and assings this label to any phraseological unit composed of at least two polysemous components, which reciprocally select a specific subsense of the other, among all the subsenses listed for a particular lexical item in the lexicon. In the case of *red herring*, for instance, the contextual presence of the morpheme *herring* assings *red* the subsense of *phony*, while the presence of *red* selects for *herring* the subsense of *issue.* Anyway, as also Strässler (1982: 32) points out, such a process can take place only after the fact, because we need to know what an idiom means to give each part its subsense. Otherwise, if the selected subsenses

were inherent to the simple lexems by themselves, we would simply have a normally compositional expression and, on top of that, we should find other instances in which *red* and *herring* mean *phony* and *issue*. Finally, as *OED* (VIII: 304) reports, this idioms has also the meaning of *soldier*: distributing this monadic sense among the idiom subparts as predicted by Weinreich's (1969) formulation becomes thorny. As regards phrase idioms and their restricted flexibility, Weinreich (1969) takes over Chomsky's (1965) generative model and introduces and *Idiom Comparison Rule* that works after the base or the transformational component: a phrase marker is built by the base component and a terminal string is generated by lexical rules and inserted in the terminal nodes of the phrase marker. Then, if this terminal string matches an entry of the *idiom list*, a device that encompasses and lists all the phrase idioms in a language, the string receives an idiomatic interpretation. Interestingly, a shortcoming shared by both Weinreich's (1969) and Katz and Postal's (1963) approach is that they cannot account for syntactically ill-formed idioms, such as *by and large*, *trip the light fantastic* or *kingdom come*, since, being syntactically anomalous, they would never be generated by the base component. Weinreich addresses this predicament by including them in the lexical-items part of the dictionary, but this would of course not explain peculiar variational preferences that some of them may display in context. In Weinreich's idea, all the particular transformational recalcitrances that are specific to a certain idiom must be listed in its lexical entry. Fraser (1970), in his fundamental contribution on idioms morphosyntactic variability from a transformational perspective, contrasts with the views mentioned so far by disposing all the idioms in a *Frozenness Hierarchy* that classifies the expressions according to their degree of flexibility. First of all, he answers the question of how idioms are represented in D-structure, that is, how they are introduced in a base P-marker generated by a set of context-free phrase structure rules. He starts from the assumption that the lexical entry of a single word like *hit* is composed of:

- a set of *insertion restrictions* that specify the syntactic context requirements: in this case, *hit* must follow a human subject noun phrase and precede a non-abstract solid or liquid object noun phrase;
- a *complex symbol* containing a set of syntactic features which dominates a phonological representation; the syntactic features specify the syntactic function of the item, [+V] for the item at hand, and other traits like [+ Voluntary Action] and

[+ Process] in this case; the phonological representation merely consists in a string of phonemes representing the word;

- a set of *semantic markers* that provide semantic information on the entry.

The lowest nodes of a base-generated P-marker are represented by complex symbols of not necessarily specified syntactic features or the phonemic representation of restricted class of formatives, like conjuctions. Insertion of lexical items happens via the comparison between a lexical entry and the structure of a P-marker. If insertion restrictions are satisfied by the environment of the P-marker and the syntactic features of the lexical item don't have different traits and values from those of the constituent-dominated complex symbol in the tree, the item is inserted in the P-marker and the two complex symbols are combined. What results is a set-union of the two complex symbols in which insertion restrictions are not present, since they are no longer relevant. Furthermore, the semantic markers formerly associated with the lexical entry are now assigned to the dominating constituent. What is the place for idioms in all of this? The deep structure representation of idioms, according to Fraser, is equivalent to that of their literal counterparts. An idiom like *pass the buck* is interpreted as a normal verbal phrase, in turn composed of a verb plus a noun phrase. Its lexical entry is likewise similar to that of a single word, the only difference being that, instead of a complex symbol dominating a phonological representation, we have a complex symbol dominating a set of complex symbols: the first specifying the syntactic function of the verb ([+V]) and its phonemic shape (*pass*), the second specifying the syntactic feature [+DET] and the phonemic string for *the* and the last one expressing the syntactic funciont [+N] and the phonological representation of *buck.* As for lexical insertion, insertion restrictions must be satisfied for the string as a whole, while the syntactic features of each component are tested against those of each component of the P-marker (a [+V], a [+DET] and a [+N] P-marker component respectively). Semantic interpretation is then assinged to the lowest constituent dominating the idiom in the tree. Supporting evidence for this treatment of idioms in D-structure comes, first of all, from the fact that idioms can undergo syntactic transformations of various kinds, just like literal phrases: *pass the buck* can be passivized in *The buck has been passed on that issue* or be converted into a gerundive nominal like in *Your passing the buck on that issue has earned you so much enmity.* Secondly, although a string like *pass the buck* is semantically ambiguous, its phonetic shape is absolutely identical whether we interpret it as an idiomatic or a literal sequence. Thus, "*by positing no difference in the syntactic structure of the idiomatic and literal*

*expressions, the same phonological rules apply without exception to both to correctly produce identical phonetic outputs"* (Fraser 1970: 26). Anyway, such an explanation leaves Fraser partially unsatisfied, since a handful of cases calls for further reflection:

a) discontinuous idioms like *bring* X *to light* or *lead* X *to a merry chase;*
b) idioms with a free possessive noun phrase like *lose* X*'s mind* or *break* X*'s heart;*
c) complex idioms like *kill the goose that lays the golden egg*, whose noun phrase contains a restrictive relative clause;
d) non-ambiguous and syntactically ill-formed idioms, like *beat around the bush* or *trip the light fantastic*
e) equating idioms with their literal counterparts in the D-structure representation, without further specification, would result in the base rules generating all the possible syntactic variants even for those idioms that permit only a restricted set of such operations (see the impossibility of action nominalization for *pass the buck*: **Your passing of the buck created much concern*) or that don't permit them at all.

A discontinuous idiom (a) like *bring* X *to light* is treated as having a complex symbol in its entry that contains other four complex symbols. The second one, X, is an empty symbol with its features constrained by insertion restrictions, which in this case would predict the verb *bring* to follow an adult human subject noun phrase and be followed by an intervening object noun phrase. Allowances are therefore made for insertion restrictions to include variables displaying specific properties. To make sure that X matches correctly with a P-marker compatible with the idiomatic string, we could also require that X possesses minimum syntactic features, in this case [+ NP]. Another solution is to conceive discontinuous idioms and verb-particle constructions in a like fashion. As though a verb-particle construction like *look up* is inserted in D-structure with its components not being separated and a movement rule that converts *look up something* into *look something up* can optionally apply, we can insert the idiom at hand in a like manner (*bring to the light something*) and then establish a mandatory movement rule that produces the usual surface order *bring something to the light*. The first solution suits also those idioms that have a free possessive noun phrase (b) if one posits a complex symbol for the variable X, which this time holds the place of a possessive determiner. For each specific case, insertion restrictions must specify whether the possessive determiner of the base P-marker must be co-referential (*lose* X*'s mind*) or not (*break* X*'s heart*) with the subject noun phrase. In an

idiom of type (c), Fraser (1970) conceives the lexical entry as encompassing the sequence "[+V; *kill*] [+DET; *the*] [+DET +WH; *that*] [+N; *goose*] [+V; *lay*] [+DET; *the*] [+ADJ; *golden*] [+N; *egg*] [+N; *goose*]". Just like the literal case, the relative clause is automatically moved after the noun to generate the surface structure. What deserves careful consideration is the noun phrase *the golden egg.* Although its literal version can be undoubtedly derived from the relative clause *The egg which is golden*, this does not apply to the idiomatic case, for which the version with the preposed adjective is the only possible one. Since we cannot assume that deriving *the golden egg* from *the egg which is golden* is obligatory or motivated by the absence of *be* in the deep structure representation, we must simply analyze this string as a non-derived determiner-adjective-noun sequence, like *the chief engineer* or *the Red Army.* Moving to point (d), Frazer proposes to analyze a non-ambiguous string like *trip the light fantastic* as a canonical verbal phrase, whose noun phrase includes a determiner-adjective-noun string, regardless of what the constituents originally meant or synchronically mean in other context. Such a representation motivates the regular phonemic shape of the expression and accounts for its eventual syntactic transformations (which are in any case impossible, being *trip the light fantastic* in the group of the most frozen idioms, see below). Similarly, an idiom like *by and large* is simply introduced in the lexicon as it is and is put into a P-marker as dominated by an adverbial constituent, since it belongs to the same class as adverbials like *certainly* or *surely.* As the author notes, the fact that no other examples of conjoined adverbials need to be introduced in the base does not matter, given that introducing such a string in the base does not have any bad consequence on the rest of the system. Crucially, in the final part of his study, Frazer (1970) intends to cope with the transformational recalcitrances that idioms display. Let's consider *pass the buck* once more: passivization (a) and gerundive nominalization (b) are allowed, but not action nominalization (c):

(3) a. *The buck has been passed on that issue.*
b. *Your passing the buck on that issue has earned you so much enmity.*
c. **Your passing of the buck created much concern.*

An *à la* Weinreich (1966) approach would enumerate, for each idiom in the lexicon, which transformations are possible and which are not: *pass the buck* would be marked as [+ Passive; + Gerundive Nominalization; - Action Nominalization]. By contrast, Fraser proposes to deal with these recalcitrances not in terms of transformations, but in terms of

*operations* on P-markers. These operations are therefore not conceived as transformations and don't correspond to them (Chomsky 1955; Fraser 1967). A conceptual switch of this kind is forced by the fact that the syntactic behavior of idioms would nullify the distinction between *governed* and *ungoverned* rules that Lakoff (1965) creates within the set of transformations. Passivization is an instance of governed rule, since its application may not occur for specific lexical items although the rest of the syntactic environment would enable it: it is the case of *suit* in the sentence *That secretary suits me fine*, which should be marked as [- Passive], since **I'm suited fine by that secretary* is not acceptable. On the flip side, particle movement, gerundive nominalization and action nominalization are ungoverned rules, given that they don't allow any exception if the relevant formal requisites are met. As we can see, this does not hold for idioms:

(4) a. **The man blew some steam off.*
b. **Your cooking your goose was stupid.*
c. **Her hitting of the sack occurred while we were visiting.*

In spite of abandoning the notion of ungoverned rule and depriving it of its meaning, Fraser decides to save it and to shift its focus on various kinds of *operations* that may be carried out on a P-marker, namely:

- *adjunction* of non-idiomatic material to the idiom, as in gerundive nominalization, with the possessive element *'s* adjoined to the NP and the gerundive marker attached to the verb (*hit the sack → John's hitting the sack*);
- *insertion* of any kind, like in some occurrences of indirect object movement (*John read the riot act to the class → John read the class the riot act*);
- *permutation* of two successive constituents, like in some examples of yes-no question (*The cat has got your tongue → Has the cat got your tongue?*) or particle movement (*Lay down the law → Lay the law down*);
- *extraction* of some idiomatic constituent outside the idiom, as in passivization (*The law was laid down by her father*);
- *reconstitution*, which entirely alters the syntactic function of an idiom. The only example is nominalization transformation: if we convert *The dad laid down the law to his son* into *The dad's laying down of the law to his son*, we have the subject

noun phrase converted into a determiner and with a possessive marker attached, *lay down* being transformed into a noun and *of* inserted after the verb.

In Fraser's proposal, all these processes can be organized in a hierarchical order, such that if an idiom can undergo a given operation, it follows that it can also undergo all the operations listed lower than the first one. The *Frozenness Hierarchy* is construed as follows:

L6 – Unrestricted
L5 – Reconstitution
L4 – Extraction
L3 – Permutation
L2 – Insertion
L1 – Adjunction
L0 – Completely Frozen

Idioms that don't have any literal counterpart belong to level 0 and don't allow variation at all, like *beat about the bush* or *trip the light fantastic*. Interestingly, no idioms can be attributed to level 6 in Fraser's opinion, since such a level would permit topicalization operations, as clefting, that cannot function for expressions composed of semantically empty and non-referential elements. The main effect of topicalization is in fact to highlight a referential element from the discourse viewpoint. As one could imagine, in a sentence like *It was the riot act that John read to me*, we could not have an idiomatic meaning, since in the idiomatic string *the riot act* does not stand for a concrete thing and does not have a concrete meaning, so it cannot be isolated and emphasized with respect to the rest of the idiom (for a critical revision of this stance, see the paragraph on Nunberg's contributions below). As regards the rest of the hierarchy, as we have told, an idiom belonging to a given level automatically belongs to all the lower ones. *Pass the buck*, for instance, is attributed to level 5, while *blow off some steam* to level one, because it only allows adjunction. As Fraser notes, an unavoidable inter-subjective difference with respect to which level a certain idiom is attributed to sometimes emerges. What remains sure, in any case, is that, whatever level we may attribute a given idiom to, it will automatically belong to the lower levels, even though a certain speaker assigns it to level 5, say, and another one to level 4. To conclude, how can the conflict with the notion of ungoverned rules be avoided? The

number of the respective level must be indicated in an idiom's lexical entry, such that no transformations can be applied to an idiom if they involve the operations prohibited at its level and at the inferior ones, with no reference to lexical exceptions of any sort.

Similar observations on the formal peculiarities of idioms are made by Van Gestel (1995) in his revision of Chomsky's (1980) *Idiom Rule* approach. According to Chomsky (1980), terminal strings of any kind, both nonidiomatic and idiomatic, are generated by base rules in a like fashion and lexical items are then inserted in $X^0$ positions. When a string is not idiomatic, it receives a nonidiomatic reading via regural semantic interpretation. Otherwise, if it matches an idiomatic string listed in the lexicon, the D-structure is reanalyzed by the intervention of an idiom rule. In the following example, such a rule deletes the syntactico-semantic features of the individual words *kick* and *bucket*, includes the object NP into the V-node and assigns the meaning "to die" to the resulting configuration:

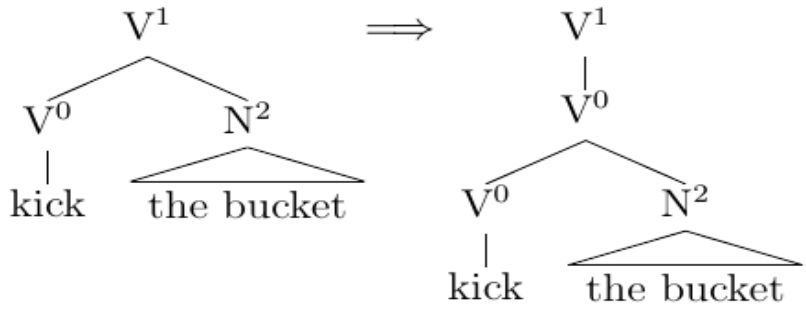



**Figure 4:** Application of the Idiom Rule (Chomsky 1980) to an idiomatic string in the P-marker

As Van Gestel (1995) remarks, not only does this rule exclude syntactically anomalous idioms, given that base rules would never generate a structure like Dutch *op en top* "all over" (lit. "up and top"), with the conjunction of heads belonging to different categories at $X^0$ level, but it also overlooks idioms like the Dutch *de huik naar de wind hangen* "to set one's sail to every wind" (lit. "to hang the cloak to the wind") that contain lexemes never appearing in isolation (in this example *huik*), cases such as *ten eigen bate* "for one's own benefit" that include a word, *ten,* resulting from the fusion of the conjunction *te* and the obsolete inflected article *den* and, lastly, idiomatic expressions in which the linear order of the single words would never be normally generated by the base rules. The last point can

be exemplified by the phenomenon of V-clusters, which arise from a verb-raising process that takes the verb of an embedded clause and lifts it to the right of the matrix verb. In a Dutch literal sentence like (5a), we can posit (5b) as the D-structure from which the V-raising takes place:

(5) a. *(dat) ik hem liet lachen*
(that) I him let laugh
'(that) I let him laugh'
b. *(dat) ik [ [hij lachen ] liet*

From (5b) to (5a), the infinitive *lachen* is raised to the right of the matrix verb *liet.* Remarkably, since this is a freely generated verbal cluster, the reversed order *lachen liet* is anyway acceptable. This does not happen for idiomatic strings like *het in Keulen horen donderen* "to be in utter bewilderment" (lit. "to hear thunder in Cologne"), which never appear with the reversed order **donderen horen*. This observation suggests that this clusters are not derivative in nature, but are already present in the deep representation of the idiom. In light of these issues, a more sensible approach for Van Gestel (1995) is to adopt a listeme approach, in which idioms are stored in the lexicon as subtrees, provided with all the idiosyncratic information that characterizes them as idiomatic. This way, words that do not occur alone like *huik* or *ten* simply fill their N positions in the subtrees, despite the fact that we would never encounter them as single words. The subtree representation of each idiom is also differentiated so as to reflect its specific degree of fixedness. Idioms like *bij voorkeur* "by preference" and *de plaat poetsen* "to cut one's stick" (lit. "to polish the plate") are represented as follows, with both the syntactic structure and the lexical items completely specified:

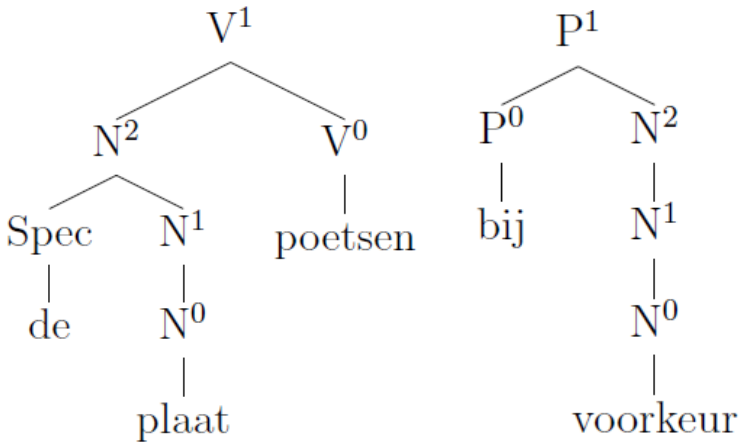



**Figure 5:** lexical subtrees representation for Dutch idioms *bij voorkeur* and *de plaat poetsen* (Van Gestel 1995: 93)

Otherwise, to represent an idiom like *koren op* X*'s molen* "nuts to someone" (lit. "wheat to someone's mill", we have to specify a free syntactic slot in the tree whose position is obligatory, while lexical choice remains open:

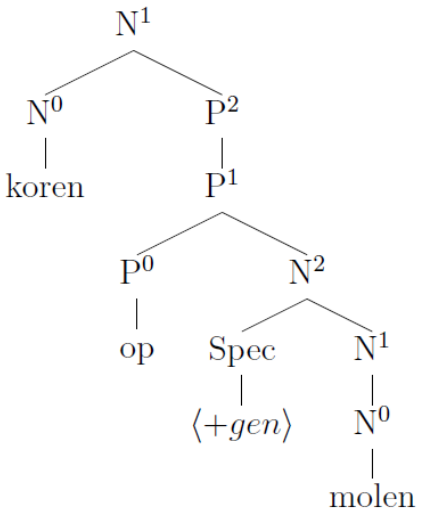



**Figure 6:** lexical subtrees representation for *koren op* X*'s molen* (Van Gestel 1995: 94)

Finally, there are idioms like *een bok schieten* "to make a blunder" (lit. "to shoot a goat") and *op termijn* "in the future" (lit. "at term") that allow insertion (*een flink bok schieten* "to make a big blunder", *op korte termijn* "at short notice"), extraposition (*de bok die hij geschoten had* "the blunder that he had made") and even morphological modification of the noun (*een bokje schieten* "to make a little blunder"). In contrast with completely frozen idioms, these represent cases of *primary idiomatization*, in which a lexical head (*schieten* and *op* respectively) selects only the head of its NP complement,

without $N^0$ level being filled beforehand:

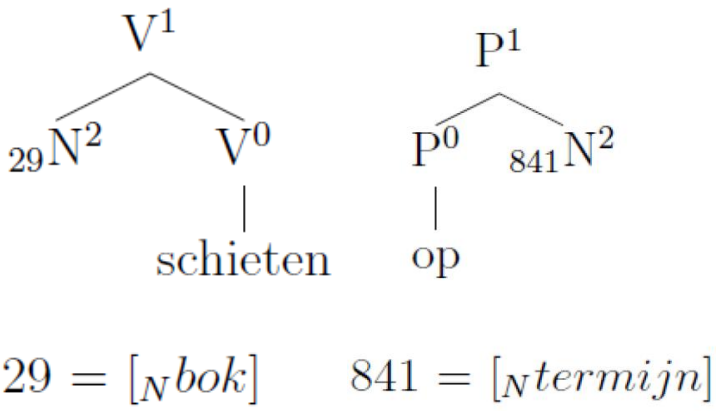



**Figure 7:** representation of *primary idiomatization* in lexical subtrees (Van Gestel 1995: 94)

This representational device is meant to show that *bok* and *termijn* are to become the heads of their NPs, but the expansion of the NP itself is left to the regular base rules. So, when it comes to the $N^1$ level, nothing would prevent an adjective from being adjoined, as in *op korte termijn.* Moreover, the noun *bok* can undergo morphological modification like the attachment of the diminutive suffix without any obstacle, since it's assigned an independent status in the syntactic structure. A further remark is made by Van Gestel about the semantics of these various idioms classes, which relates to what Nunberg and colleagues (Nunberg 1978; Wasow et al. 1984; Nunberg et al. 1994) affirm about the semantic analysability of such expressions (see the following paragraph): the proposed representation of frozen idioms like *de plaat poetsen* clearly predicts that the N and V don't have any semantics of their own and that the NP receives no theta role, while the isolated status of *bok* in the X-bar representation of *een bok schieten* highlights that it has its own referentiality and its own contribution to the meaning of the whole idiom.

The subtrees of the listemes are inserted in the syntactic structure via *en bloc insertion*: all in all, if the structure of a subtree matches the syntactic structure of a sentence that is being generated, the idiom is inserted. To cope with the fact that the base rules would never spontaneously generate syntactically anomalous structures, Van Gestel (1995) proposes an optional application of the X-bar rules that provide the structure for a sentence. Just like, during structure generation, X-bar core rules in (6) can freely interrupt to enable the application of the more peripheral rules in (7), among which adverbial and adjectival insertion, X-bar rules can stop at an adverbial $X^2$ to permit the insertion of an adverbial idiom like *by and large*, without any further expansion of the node nor any

specification of the internal structure of the expression, or at a $V^1$, so that a $V^1$ idiom like *kick the bucket* can enter the structure.

(6) a. $X^3 \rightarrow N^2\ X^2$
b. $X^2 \rightarrow [\text{Spec}\ X^1]\ X^1$
c. $X^1 \rightarrow Y^n,\ X^0$

(7) a. $X^3 \rightarrow Adv^2\ X^3$
b. $X^2 \rightarrow -V^3\ X^2$
c. $N^2 \rightarrow N^2 - V^3$
d. $N^1 \rightarrow A^3\ N^1$

Focusing on Italian, Bianchi (1993) classifies a set of 28 verbal idioms into three classes according to the kind of syntactic movement they tolerate:

| | **I** | **II** | **III** | **Arguments** |
|---|---|---|---|---|
| **Left dislocation** | + | + | + | + |
| **Non-quantificational movement** | - | + | + | + |
| **(Restricted) quantificational movement** | - | - | + | + |
| **Bare wh-phrase** | - | - | - | + |
| **Clefting** | - | - | - | + |
| ***Tough*-movement** | - | - | - | + |

**Table 1:** the classification of Italian idioms advanced by Bianchi (1993)

Idioms in group I include *tagliare la corda* ('to take French leave', lit. 'to cut the rope'), *ficcare il naso* ('to poke nose into'), *alzare il gomito* ('to bend one's elbow', lit. 'to raise the elbow'), *sbarcare il lunario* ('to make ends meet', lit. something like 'to unload the almanac'). Group II includes *fare gli onori di casa* ('to do the honors', lit. 'to do the home honors'), *fare giustizia* ('to make justice'), *dare una lezione* ('to punish someone', lit. 'to give a lesson to someone'), *dare il buon/cattivo esempio* ('to set a good/bad example') and *ingoiare un boccone amaro* ('to swallow a bitter pill', lit. 'to swallow a bitter bite'). Group III includes *dedicare tempo* ('to devote time to someone/something'), *prestare attenzione* ('to pay attention', lit. 'to lend attention'), *rendere onori* ('to pay homage', lit. 'to render honors'), *fare debiti* ('to get into debt', lit. 'to make debts'), *fare progressi* ('to make headway', lit. 'to make progresses') and *prendere l'iniziativa* ('to take the initiative').

In the fourth column true referential arguments are added, namely those arguments that have a referential thematic role like Agent, Patient or Experiencer and indicate a participant in the event or situation expressed by the predicate (Rizzi 1990). While all kinds of idioms and arguments accept left dislocation, specific subsets of idioms behave differently with respect to quantificational and non-quantificational movement (Lasnik & Stowell 1991). This distinction concerns A' movements, in which, according to the traditional definition in Government and Binding theory (Chomsky 1981), an element is moved to a position that is not assigned a theta role. In quantificational movement, the moved element is a Quantifier Phrase that fulfills the following requirement (Bianchi 1993: 357):

*A true Quantifier Phrase is composed of a quantifier Q and a nominal term T defining a range R that Q quantifies over, such that R is a possibly non-singleton set.*

In this case, the bound trace works as a variable. This kind of movement, exemplified in the sentences below, shows the Weak Crossover Effect (WCO; Lasnik & Stowell 1991): when a trace is bound by an interrogative or a relative pronoun in a restrictive clause, it cannot be coindexed with a pronoun on its left:

(8) a.**Who*$_i$ *did his*$_i$ *neighbors killed* $t_i$*?*
b. **Every teacher*$_i$ *that his*$_i$ *pupils love* $t_i$ *is satisfied.*

WCO is not observed with non-quantificational movement, like in the *easy-to-please* construction, in topicalization or in non-restrictive relative clauses, where the A' operator receives his reference from a fixed antecedent and the bound trace has not range over a domain of quantification:

(9) a. *Jim*$_i$ *is hard for his*$_i$ *mum to keep under control* $t_i$.
b. *Jim*$_i$*, I see his*$_i$ *mum keeps under control* $t_i$.
c. *Jim*$_i$*, whom*$_i$ *his*$_i$ *mum tries to keep under control* $t_i$*, is a real nuisance.*

Although non-quantificational movement is theoretically distinct from A movement, Bianchi (1993) assimilates them for the sake of the current argumentation, since the subclass of idioms accepting non-quantificational movement permits A movement as well.

In any case, what both kinds of operations share is that the moved element is not necessarily quantified. Foremost, idiomatic expressions in group II and III tolerate passivization (10) and topicalization (11), differently from group I:

(10) a. **La corda è stata tagliata da Gianni.* (Group I)
b. *Gli onori di casa sono stati fatti da Lucia.* (Group II)
c. *L'iniziativa è stata presa dal comitato.* (Group III)

(11) a.**IL NASO ficca negli affari altrui.* (I)
b. *UNA BELLA LEZIONE dovresti dare a quel mascalzone.* (II)
c. *GRANDISSIMI ONORI si dovrebbero rendere a questo eroe.* (III)

Bianchi observes that the impossibility of passivization in (10a) cannot be accounted for by a violation of adjacency requirements for all the idiom chunks in the Logical Form (LF), as is traditionally assumed (Chomsky 1992: 30). Recalling what Belletti (1990) affirms about Italian verb movement, the verb in S-structure raises to Agreement-Subject, the highest functional head in the sentence. Such a positional change allows adverbs to intervene between the verb and the VP complements, in literal expressions as much as in idioms (12):

(12) a. *Gianni non scrive più romanzi.*
b. *Gianni non taglia più la corda.*

In brief, in both idioms and literals the verb raises, but the object NP stays within the VP, so the chunks are not generated in adjacent positions. Passivization without preverbal movement of the deep object gives rise to the same grammaticality judgments of (10):

(13) a. **È stata tagliata la corda da Gianni.* (I)
b. ?*Sono stati già fatti gli onori di casa.* (II)
c. *Finalmente è stata presa l'iniziativa di organizzare una riunione.* (III)

This is because also the sentences in (13) are instances of A movement: the expletive in the preverbal position lacks semantic content and must be eliminated in LF according to the Full Interpretation Principle (Chomsky 1986; 1991). So, the postverbal subject that forms a chain with it, moves to preverbal position and replaces it. Consequently, we can

affirm that the LF of the (13) sentences more or less corresponds to the S-structure of the (10) sentences. To sum up, Bianchi (1993: 366) states that what separates idioms tolerating A and non-quantificational movement from those not allowing it is the presence of an invariable determiner (**tagliare una corda, *tagliare questa corda* vs. *prendere un'iniziativa, prendere questa iniziativa*), the impossibility of free modification (**tagliare la agognata corda* vs. *dare una bella lezione*) and the total semantic noncompositionality. Crucially, lack or assignment of theta roles is seen as the syntactic correlate of this difference between more and less compositional idioms. If idioms of type I are totally noncompositional, their object does not possess a semantics on its own and it can be hypothesized that it doesn't even receive any thematic role by the verb. Being semantically empty, it is deleted from the LF according to the Full Interpretation Principle. This can happen either via the application of an idiom rule that leaves an empty category in its place (Chomsky 1991) or via $V^1$ reanalysis, that unites the idiomatic NP and the verbal trace into a $V^0$ complex verb. In either way, the idiomatic NP must remain in the base position to be deleted at LF without any repercussion on the sentential structure. Otherwise, as in (13a), we would have a chain between a head element (the expletive) and a deleted position (the postverbal idiomatic NP), which would make the head element uninterpretable. The fact that idiomatic NPs of type II and III do receive theta roles is first of all motivated by the observation that idiom chunks somehow contribute to the overall idiom meaning[2]. On top of that, it is corroborated by some examples coming from French (Bianchi 1993: 367), where idioms like *tirer parti de* ('to take advantage of'), *rendre la justice* ('to dispense justice') or *porter secours* ('to bring help') are grammatical in the passive form (15) if and only if the corresponding impersonal passive (14) is acceptable:

(14) a. *Il a été tire parti de cette affaire.*
b. *Il a été rendu justice dans cette ville.*
c. *Il a été porté secours aux victimes de cet accident.*

(15) a. *Parti peut être tiré de cette affaire.*
b. *Justice a été rendue dans cette ville.*
c. ? *Secours a été porté aux victimes de cet accident.*

[2] With respect to the MWEs classification we sketched above, some idioms of type III such as *prestare attenzione*, *rendere onori* e *fare progressi* could even be classified as LVCs, with the noun element giving the most notable contribution to the overall meaning and the verb being almost semantically dummy.

It must be noted, at this point, that subject inversion like in (14) is allowed in French only for unaccusative verbs and indefinite NPs. According to Belletti (1988), this is due to the fact that postverbal subjects receive inherent partitive case from the verb, while nominative case is assigned to the expletive pronoun and definite NPs cannot appear in the partitive. Since inherent case assignment is just possible for theta-marked positions (Chomsky 1986), we can conclude that idiomatic NPs in (14) possess a thematic role.

Finally, a subset of idioms that can undergo non-quantificational movement, namely idioms of type III, are also liable to quantificational movement, including restrictive relativization (16), amount interrogation (17) and *what*-interrogation (18):

(16) a. **La corda che ha tagliato…* (I)
b. **La giustizia che ha fatto…* (II)
c. *L'iniziativa che hai preso è discutibile.* (III)

(17) a. **Quanta corda ha tagliato?* (I)
b. **Quanta giustizia hanno fatto?* (II)
c. *Quante iniziative ha preso senza consultarti?* (III)

(18) a. **Che corda ha tagliato?* (I)
b. ?**Che attenzione ti hanno dedicato?* (II)
c. *Che iniziativa ha preso?* (III)

Most notably, all the idioms of this type, together with some idioms belonging to the second group (*dare una lezione, dare il buono/cattivo esempio, ingoiare un boccone amaro*), display a variable determiner (e.g. *l'iniziativa* vs. *quante iniziative*). This permits actual quantification and NP modification to take place, while the invariable determiner of some idiomatic NPs of type I and II (e.g. *la corda* vs. **quanta corda*) *"is void of quantificational force"* (Bianchi 1993: 376). A related phenomenon is the total absence of determiner in some idioms like *fare giustizia*. In the realm of referential arguments, nouns with a cumulative reference (Link 1983), i.e. mass or plural nouns, can have a zero determiner represented by a phonetically empty category $D^0$ that receives an existential interpretation at LF (Longobardi 1992). In the case of *fare giustizia* and other similar idioms, the lack of an overt determiner co-occurs with a singular count noun. Bianchi (1993: 376) proposes that these idioms do not display any determiner at all and that this assumption explains why these expressions do not satisfy the cumulative reference requirement. The presence of a fixed determiner and the impossibility of modification in

those idioms that cannot undergo quantificational movement is motivated by the lack of an open position in the element to be moved. Higginbotham (1987) asserts that predicative categories (e.g. modifiers) have an open position in their thematic grid, which coincides with a variable over the Universe of Discourse. Noun possess an open slot too and indicate a set of referents. In modification, the open position of a modifier is theta-identified with the open position of the nominal projection, resulting in an intersection of the denotations of the two elements. Determiners, on the other hand, bind the open position of the nominal projection and generate a saturated category that cannot function as a predicate anymore. Both processes cannot apply to idioms of type I and II, since the nominal projections of these idiomatic expressions lack an open position.

In the end, *tough*-movement (19), clefting (20) and bare wh-phrase questions (21) are only allowed for true referential arguments. A provisional explanation provided by the author is that the moved element must have a referential index as a well-formedness condition (Bianchi 1993: 382).

(19) a. **La corda è difficile da tagliare.* (I)
b. **(La) giustizia è difficile da fare.* (II)
c. ?? *L'iniziativa è difficile da prendere in queste circostanze.* (III)

(20) a. **È la corda che ha tagliato.* (I)
b. ** È giustizia che hanno fatto.* (II)
c. ** È l'iniziativa che ha preso senza consultarmi* (III)

(21) a.**Che cosa ha tagliato? – La corda.* (I)
b. **Che cosa hanno fatto? – Giustizia.* (II)
c. **Che cosa ha preso? – L'iniziativa.* (III)

### *1.4.3. Challenges to idiomatic noncompositionality: the typologies of Nunberg and colleagues (1984; 1994) and Cacciari and Glucksberg (1991)*

Scholars adhering to transformational frameworks largely resort to idioms in their arguments for the existence of transformations (Culicover 1976; Keyser & Postal 1976; Chomsky 1980: 149 ff.). If the four sentences in (21) were directly generated in their surface structure and if an idiom were merely an arbitrary pairing of form and meaning, each of the four instances should enter the lexicon as a different idiom:

(21) a. *Mary pulled some strings to get that job.*

b. *Some strings were pulled by Mary to get that job.*

c. *Some strings seem to have been pulled by Mary to get that job.*

d. *Some strings are believed to have been pulled by Mary to get that job.*

On top of that, idiomatic chunks can be indefinitely separated as in (d) or in even more complex structures, ending up with an almost infinite set of variants for each idiom to be listed. Contrariwise, this grammatical machinery could be conceived more economically if one posited a deep structure in which idioms are collocated in their contiguous form (a) and transformational operations such as passivization or raising that derive the surface forms (b-d). As supporters of alternative conceptions like Lexical Functional Grammar (Bresnan 1981) or Generalized Phrase Structure Grammar (Gazdar 1982) sustain, the assumption that the only way for a grammar to capture active/passive and raised/non-raised relations is through transformation appears questionable: Bresnan does it by means of lexical redundacy rules and Gazdar with metarules. More importantly, if the received equation between idiomaticity and noncompositionality turns out to be unfounded, as Nunberg (1978) already claims, one could simply expect idioms to normally have their passive counterparts, given that actives and corresponding passives possess the same predicate-argument structure (Wasow et al. 1984). More in general, recognizing that idioms are indeed composed of chunks, each of them separately contributing to the overall meaning, can also semantically motivate their well known transformational recalcitrancies (Chafe 1968; Wasow et al. 1984; Nunberg et al. 1994; Cacciari & Glucksberg 1991; Bianchi 1993). Chomsky (1980) observes that while *take care of* allows the passive (22a), the idiomatic meaning is not preserved with the tough-movement (22b).

(22) a. *Excellent care was taken of the orphans.*

b. *Excellent care is hard to take of the orphans.*

For sentences of this kind, his paradigm predicts the application of idiom rules in D-structure. In (22a), the idiomatic object is contiguously generated with the rest of the expression and then moved to the subject position once the entire sequence has received an idiomatic interpretation from the just mentioned rules. On the other hand, (22b) is intepreted as an instance of deletion rather than movement: the subject is not moved but

generated in place and the string can consequently receive no idiomatic reading. These variational restrictions are therefore explained with idioms differently behaving with movement and deletion operations (Wasow et al. 1984). Actually, tough-construction examples for idioms are attested (Berman 1974):

(23) a. *That nerve is easy to touch.*

b. *Sometimes the bullet is hard to bite.*

To explain why some idioms accept this variation, Wasow et al. (1984) sustain that the discriminating factor does not reside in the movement/deletion distiction, but in the opposition between idiomatic chunks that are arguments of their predicates and those which aren't. Adjectives like *easy* and *hard* in (10) are thought of as binary predicates requiring their surface subject as one of their argument: to do so, the subject is supposed to effectively behave as an argument, that is, it has to provide its own independent meaning to that of the whole expression. Since sentences like (23a) and (23b) do sound acceptable to native speakers, in the end we must admit that some idioms are, at least *sensu lato*, compositional.

This position, first proposed by Nunberg (1978) and Wasow et al. (1984), affirms that, while a speaker cannot grasp the meaning of a given idiom like *pull strings* by hearing it isolation, but he needs encountering it in an informative context (e.g. *John was able to pull strings to get the job, since he had a lot of contacts in the industry*), once he has learnt it, he can *"establish correspondences between the parts of the structured denotation of the expression […] and the parts of the idiom"* (Nunberg et al. 1994: 496). In this case, once we have learnt to associate *pull strings* with the meaning of "exploit personal connections", we can interpret *pull* as metaphorically standing for the act of exploiting and *strings* as metaphorically standing for the personal connections the subject has. On the contrary, we cannot detect this mapping between pieces of the idiomatic referent and pieces of the idiomatic string for cases like *kick the bucket*: the meaning of "die" cannot be sensibily decomposed into elements that roughly correspond to the meanings of *kick* and *bucket;* likewise, for *shoot the breeze* we cannot find a mapping of any sort between the meaning of "conversate idly" and the meanings of *shoot* and *breeze*. While Wasow et al. (1984) define the first class of idioms as *compositional,* they subsequently redimension such a statement, preferring the labels *idiomatically combining expressions* for decomposable idioms like *pull strings* or *pop the question* and *idiomatic phrases* for non-decomposable

idioms like *kick the bucket* or *trip the light fantastic* (Nunberg et al. 1994). Speaking in terms of strict sense compositionality appears *de facto* imprecise and questionable: while a perfectly compositional sentence can be understood by a native speaker even though he hasn't witnessed it before, the semantic decomposition of an idiom takes place only after the fact, that is, after a speaker has stored its figural interpretation. In order to avoid terminological confusion, researchers generally use the term *semantic analyzability* or *decomposability* in place of *compositionality* with respect to idioms (Fazly et al. 2009). More precisely, the class of decomposable idioms is further divided by Nunberg (1978) into *normally* and *abnormally* decomposable ones. *Pop the question* is an example of normally decomposable idiom, since there is a direct mapping between *pop* and *question* on the one hand and the meanings "suddenly ask" and "marriage proposal" on the other. In *spill the beans* and *pull strings* the mapping is, contrariwise, metaphorical and not direct: we accordingly label them as abnormally decomposable idioms. In any case, even if we can establish a connection between idiom chunks and pieces of an idiomatic meaning, we cannot explain why it is *beans*, for instance, that are spilled in revealing a secret and not some other kind of vegetable or why it is *strings* that are pulled in exploiting one's connections and not, say, *cords* or *ropes.* In other words, we cannot explain why the idiomatic reference is metaphorically expressed by those exact words and not others. This is due to the fact that *decomposability* does not go hand in hand with *transparency*: *spill the beans* and *pull strings* are decomposable but not transparent; by contrast, an English speaker could easily guess why *saw logs* means "sleep", since the two actions produce a similar noise, but he could not assign parts of the idiomatic meaning to the two component words. We must then ask ourselves whether the possibility of semantically motivating the internal parts of an idiom means that we can use them in isolation with their idiomatic meaning. The answer is, generally speaking, negative: *spill* means "divulging" only when it co-occurs with *beans* and *beans* means "secret" only when it's used with *spill*. To motivate such an interpretative constraint, Nunberg et al. (1994: 505) define the relation between the verb and the NP in an idiomatic combination as a sort of semantic dependency. *Spill the beans* consists of an idiom in which a literal "spilling-the-beans" meaning is paired with a figurative "divulging-a-secret" meaning. When *beans* occurs without *spill*, the literal meaning to which the idiomatic meaning is mapped cannot be fully realized and, as a consequence, no figural meaning can be accessed and distributed among its parts. This can be seen as an extreme case of selectional restrictions, in which the semantic domain of both the idiomatic noun and the idiomatic verb are singleton sets

(Wasow et al. 1984: 93). English language, however, also presents cases of names and verbs with non-singleton domains. It is the case of *idioms families*, in which the same verb can appear with different NPs to form distinct, but semantically related, expressions (24) and vice versa (25):

(24) *hit the hay/sack; lose one's mind/marbles; open the floodgates/sluice gates/gates; drop a bomb/bombshell/brick*
(25) *keep/lose/blow one's cool; step/tread on someone's toes; beat/whale the tar out of someone; stop/turn on a dime*

As Nunberg et al. (1994: 504-5) underline, traditional accounts that don't see idioms as semantically analyzable (see the preceding paragraph) fail to capture such generalizations. A still different case in which the idiomatic noun has a non-singleton semantic domain is represented by *pull strings.* Since the following examples are grammatical:

(26) a. *Pat pulled strings that Chris had not access to.*
b. *The strings that Pat pulled helped Chris get the job.*

we must infer that idiomatic *strings* belongs not only to the domain of *pull,* but also to the domain of the intensions of *have access to* and *help Chris get the job*. Nevertheless, this would force us to accept a figurative reading also in cases like:

(27) a. *Chris had no access to strings.*
b. *Strings helped Chris get the job.*

Wasow et al. (1984: 94) reach a compromise by stating that such a nonliteral interpretation is influenced by specific condition of use and can take place only if the entire idiom has previously been cited in the same discourse, as in the following passage:

(28) *Pat and Chris graduated from law school together with roughly equal records. Pat's uncle is a state senator, and he pulled strings to get Pat a clerkship with a state supreme court justice. Chris, in contrast, didn't have access to any strings, and ended up hanging out a shingle.* (Wasow et al. 1984: 94)

If co-occurrence of all the idiom chunks is required for the figural meaning to be activated, once these conditions have been fulfilled and the idiomatic meaning has been distributed among the components, they can be used in isolation.

Most importantly, the separation between idiomatically combining expressions and phrasal idioms appears to be reflected in their different syntactic behavior (Wasow et al. 1984: 90 ff.; Nunberg et al. 1994: 499 ff.), in accordance with Chafe's (1968) and Newmeyer's (1974) observations that the transformational restrictions of idioms can be predicted at least in part on the basis of their meaning.

First of all, idiomatic combinations can undergo adjectival insertion or relativization so that the modification involves just a specific piece of the idiomatic reference:

(29) a. *Sam smoked for years, but then she kicked the unhealthy habit.*

b. *After days of embarrassment, I finally broke the ice that had formed between us.*

In (29a), it is the smoking habit to be unhealthy and not the whole fact of kicking it. In (29b) the relative proposition does not refer to the entire act of relaxing a tense situation, but just to the awkward situation itself, which is metaphorically represented by the ice. This is what Ernst (1981) calls *internal modification.* As we already stated, if we modify an idiomatic phrase as in *kick the proverbial bucket*, the adjective does not refer to a specific part of the idiomatic meaning "die", since we cannot divide it among the constituent words, but it modifies the idiom as a whole, resulting in a sort of metalinguistic comment on the expression itself. For idiomatic phrases we therefore talk about *external modification* (Ernst 1981). It's worth observing, anyway, that although *kick the bucket* is a non-analyzable idiomatic phrase, the verb *kick* contributes with its actional features to the syntactic restrictions of the whole expression (Nunberg 1978; Wasow et al. 1984: 92; Cacciari & Glucksberg 1991: 233). As a punctual verb, *kick* cannot occur with durative temporal adverbials like *for X time* (30a) unless the context allows an iterative reading (30b):

(30) a. ? *Mary kicked the ball for ten minutes.*

b. *The bullies kicked poor Jim in the stomach for at least ten seconds.*

On the flip side, the achievement *die* can occur with this type of adverbial to indicate the length of the time span after which the change of state implied by the lexeme takes

place (31):

(31) *The old man lay in the bed dying for a week.*

Quite surprisingly, even though *kick* does not preserve its individual meaning when it appears in the idiomatic string *kick the bucket*, it nonetheless keeps its individual incompatibility with durative adverbials and a sentence like (32) would consequently look ungrammatical for a native speaker:

(32) **The old man lay in the bed kicking the bucket for a week.*

It goes without saying that we cannot resort to an iterative interpretation to make (32) sound acceptable, since the act of dying cannot, of course, be reiterated.

Another kind of internal modification that can occur with idiomatic combinations and not with idiomatic phrases is quantification. (33a) does not mean "to reveal a secret twice", but "to reveal a couple of secrets", while (33b) is not acceptable, given that *breeze* is not homomorphically mapped with part of the meaning "to chat idly":

(33) a. *spill a couple of beans*
b. **shoot a couple of breezes*

Idiomatic combinations also allow topicalization, which typically requires the topicalized constituent to have its own independent meaning so that it can be given discourse prominence (34):

(34) a. *The question, he might pop when you least expect it.*
b. *These beans, he might spill whenever he gets distracted.*
c. **The logs, I want to saw as soon as I get home.*
d. **The bucket, he might kick if he doesn't watch out.*

Passivization is another discriminating factor: since only referential nouns can appear as surface subjects in a passive sentence (Gibbs & Nayak 1989), we can accept this operation only with idiomatic combinations (35a) and not with idiomatic phrases (35b):

(35) a. *Advantage was always being taken of him, since he was too generous.*

b. **The hay was hit by all the guests in the inn.*

It is well established that, for VP ellipsis to occur, the antecedent of the missing element must coincide with a semantic unit (Sag 1976). The acceptability of (36) supports the theory of Nunberg and colleagues about idiomatic combinations:

(36) a. *You think advantage is always being taken of you, but in fact it isn't.*

b. *My goose is cooked, but yours isn't.*

As regards anaphoric reference, Bresnan (1982: 49) denies its applicability to idioms. In fact, just like the other syntactic operations mentioned so far, evidence shows that it is possible for idiomatic combinations (37a) but not for non-decomposable strings (37b):

(37) a. *I thought Mary would break the ice, but it was John to break it eventually.*

b. **I was about to hit the hay, but then I decided not to hit it and I watched a movie instead.*

The same reasoning applies to Italian. Cinque (1990: 162) affirms that idiomatic components, like measure phrases, cannot be antecedents for pronouns except for left dislocation, since are both nonreferential, but human-elicited judgments actually regard all the sentences in (38) as acceptable:

(38) a. *Maria non ha mai pesato 70 chili ed anche suo figlio non li ha mai pesati.*

'Maria has never weighed 70 kilos and even her son has never weighed them'.

b. *Se Andreotti non farà giustizia, Craxi la farà.*

'If Andreotti will not do justice, Craxi will do it'.

Once more, the only obstacle seems to be the presence of an idiomatic phrase (39):

(39) **Gianni ha tagliato la corda, ma Paolo non l'ha tagliata.*

'Gianni has taken French leave, but Paolo hasn't taken it'.

Aside from semantic and syntactic criteria, the functional typology of idioms proposed

by Cacciari and Glucksberg (1991: 228 ff.) and Glucksberg (1993; 2001) is also grounded on differences in lexical and discourse productivity. The four classes they single out are:

1. *non-analyzable idioms*, type *N* (e.g. *by and large*, *spic and span*)
2. *analyzable-opaque idioms*, type *AO* (e.g. *kick the bucket*)
3. *analyzable-transparent idioms*, type *AT* (e.g. *break the ice*)
4. *quasi-metaphorical idioms*, type *M* (e.g. *give up the ship*, *carry coals to Newcastle*)

Both N and AO idioms can roughly be included in Nunberg et al. (1994) idiomatic phrases category. Theoretically speaking, idioms of type N should be nonproductive from the semantic, syntactic, lexical and discourse viewpoint. Actually, if lexical substitution of any component or meaning-preserving syntactic operations are undoubtedly impossible, in light of what we have already observed for idiomatic phrases, some minimal semantic and discourse productivity is permitted. Taking *by and large* as an example, Cacciari and Glucksberg (1991: 231) notice that despite its general non-analyzability we can somehow relate the word *large* to the overall meaning of "generally". As a consequence, if the idiom has already been used in a discourse, the semantics that is attributed to these components remains available for further elaboration (40):

(40) A: *By and large, people are well-off these days.*
B: *By and not-so-large! Have you seen the figures on homelessness in America?*

B speaker's response plays on the semantics of *large* and produces the form *by and not-so-large* in which the scope of the negation is confined to a single element of the string. Because, as Cruse (1986) states, negation or adjectival modification cannot have an empty element within their scope, we must assume that at least some minimal semantic information can be attributed to the single *large* component and that *by and large* is not *stricto sensu* a type N. Given that such internal modification is allowed even for a traditionally non-decomposable idiom, Cacciari and Glucksberg (1991: 232) go so far as to say that *"indeed, pure type N idioms may not exist at all".* In a similar spirit, Glucksberg (2001) sustains that specifying the syntactic idiosyncrasies of an idiom aprioristically and outside a context does not make any sense. According to his view, almost any operation can take place, on condition that it respects the semantics of the constituents and displays a plausible communicative intention: while passivization is not usually tolerated for *bury the*

*hatchet*, it may sound perfectly grammatical if enough context is provided, as in *"After years of murderous warfare, the hatchet was finally buried once and for all"* (Glucksberg 2001: 86). Similarly to N idioms, AO type idioms don't appear semantically decomposable, but their component elements nonetheless contribute to a greater extent to the semantic and syntactic behavior of the whole expression, as we have seen for the incompatibility of *kick the bucket* with durative adverbials. Lexical flexibility is likewise very reduced and generally allowed only under specific discourse circumstances. We have already reported that Moon (1998) finds *kick the can* and *kick the pail* as lexical variants in her corpus analysis. Though such alternative forms would still be associated with the "die" meaning (Gibbs et al. 1989), the choice of alternative words would normally look unjustified for a native speaker and probably interpreted as a slip of the tongue, since the internal parts of the string don't have a semantic consistency on their own. An exception could be represented by some restricted conversational contexts (41) in which the interlocutor might want to lessen a previous statement:

(41) A: *Did the old man kick the bucket last night?*
B: *Nah, he barely nudged it.*

While AT idioms substantially correspond to Nunberg et al. (1994) idiomatically combining expressions, M type idioms constitute a noteworthy class, since their literal referent corresponds to an ideal or prototypical instance of their idiomatic referent. *Be two peas in a pod*, for instance, represents both a prototypical exemplar of a situation of total resemblance between two elements and a phrase that can refer to any kind of perfect resemblance. So do *give up the ship* and *carry coals to Newcastle*: they denotate ideal scenarios of "surrendering" and "bringing something to a place that already has it in abundance", but they can also figuratively apply to every situation of the same two sorts. M idioms are thus interesting in that they partly retain the metaphorical status that most idioms originally had:

> "[…] the majority of idioms began their lives as metaphors; and syncronically, transitional cases, which are idioms for some and metaphors for others, are not uncommon" (Cruse 1986: 44)

Glucksberg and Keysar (1990) claim that metaphors like *My job is a jail* are not to be

analyzed as implicit similes, as it's traditionally assumed, but as class-inclusion assertions: a referent (e.g. *my job*) is mapped to a diagnostic or evalutative category (e.g. unpleasant situations that constrain one), which is prototipically represented by the metaphor vehicle, in this case *jail.* The same function is performed by quasi-metaphorical idioms. In saying something like (42):

(42) *Those twins are really two peas in a pod.*

we are including the twins at hand in the category of perfectly similar-looking entities by making reference to a prototypical instance of perfect resemblace, namely that of two peas in a pod. Interestingly, also metonymical expressions like *bury the hatchet* or the Italian *mettersi le mani nei capelli* ("to despair", lit. "to put one's hand in one's hair") belong to this category, because, from representing just a part of an archetypical situation of peace-making or desperation, they pass to denote the situation as a whole. From a variational point of view, they exhibit similar features to *AO* and *AT* class.

As Cacciari and Glucksberg (1991) highlight, the relevance of this distinction between analyzable and non-analyzable idioms is supported by psycholinguistic findings (Cacciari & Tabossi 1988; Gibbs et al. 1989) that see idiomatic strings as processed both semantically and syntactically. Accordingly, the processing cost of such strings increases if the results of both semantic and syntactic analysis are not congruent with each other, i.e. when the sequence is not decomposable and its single words don't seem to have any well-defined relation to its overall meaning.

This last consideration leads us to another major point in our introductory overview on idiomatic expressions: theoretical reflection on idiom semantics, syntax and typology goes hand in hand with psycholinguistic research on the mechanisms underlying idiom comprehension and production, with all the mutual suggestions, confirmations and denials that follow. We therefore consider beneficial to present a concise review on the major existing hypotheses on idiom processing.

### *1.4.4. Psycholinguistic models of idiom processing*

The first notable study to address idiom comprehension is conducted by Bobrow and Bell (1973). What they want to test is whether effectively idioms are processed as words,

differently from literal expressions, wherein the word meanings are individually retrieved and then related to build the sentence meaning (Quillian 1968). Employing a perceptual set paradigm (Marshall 1965), they have subjects read a set of literal or idiomatic sentences (set sentences) followed by an ambiguous sentence (test sentences), for which they have to mark whether the idiomatic or the literal sense is perceived first. Taking as a baseline the proportion of subjects that perceive a given ambiguous sentence as idiomatic without prior exposure to a set, the idiomatic set is shown to increase the percentage of subjects that perceive the idiomatic meaning in the test sentence first with respect to the baseline, while the literal set lowers the percentage to the baseline level. Since the set sentences do not exhibit any cues which could foster either of the two readings of the test sentences, the different results are ascribed by the two scholars to differences in processing. On the basis of their results, they endorse an *Idiom List Hypothesis*, which we can include among the so-called *Lexical look-up* theories and which predicts that idioms are represented as semantically empty long words in a mental list separated from the mental lexicon. When a string is being processed, the literal meanings of its words are first retrieved and combined. If the speaker doesn't wind up with a feasible interpretation, the idiom list is then checked to find a stored expression that matches the given string (Cacciari & Glucksberg 1991: 218). *Stage models* like this, that see literal and figurative processing as two successive phases, derive from a sequential interpretation of Searle's (1975: 114) treatment of figurative sentences. Searle (ibid.) claims that expressions like metaphors and idioms are essentially defective in that they are literally nonsensical and violate speech acts rules and conversational principles. Therefore, if a speaker recognizes a linguistic string as defective, he/she must then find an utterance meaning which is different from the literal one. However, plenty of evidence proves this prediction wrong: subjects presented with stories biasing a literal or idiomatic interpretation of the ambiguous idioms that end them assign faster interpretation to idiomatic rather than literal targets (Ortony et al. 1978). Gibbs (1980) finds that participants are faster in judging the appropriateness of the figural paraphrase of a given idiom when it is predeced by an idiomatically biasing context than they are in judging a literal paraphrase after exposure to a literally biasing context. Finally, speakers appear to understand familar idioms at least as quickly as their literal counterparts (Swinney & Cutler 1979; Gibbs 1980).

Swinney and Cutler (1979) measure reaction times in a phrase classification task in which subjects are presented with idiomatic, literal and nonsense strings and have to decide whether a given string is an acceptable English expression or not. Participants are found to

respond faster to idiomatic strings not only despite different transitional probabilities among the words in the stimuli or different degrees of flexibility among the chosen idioms, but also despite the different level of awareness each subject has about the presence of idioms in the stimuli set. According to the authors, the fact that subjects classify idiom faster even if they are not aware of their presence in the set weakens the possibility of a specialized idiom processing mode (Bobrow & Bell 1973) and instead supports a *Lexical Representation Hypothesis*, which sees idioms as stored and accessed in the mental lexicon together with any other word. Any assumption about the existence of a separate idiom list is therefore put aside. Moreover, faster response times to idioms suggest that when a speaker encounters the first word a given string, the retrieval of the idiomatic meaning and the computation of the compositional meaning are initiated in parallel. Since word recognition usually takes less time than phrase comprehension (Cacciari & Glucksberg 1991: 218), the figural meaning becomes available before the literal one. A further refinement of this idiom-superiority vision is advanced by Gibbs (1980) in the light of his findings that the idiomatic paraphrase of an ambiguous string, when preceded by an idiomatic context, is judged as appropriate more rapidly than the literal paraphrase of the same expression after exposure to a literal context. He suggests that the main discriminating factor that best explains these results could not be the difference between literal and idiomatic language, but between conventional and unconventional usage of the same expression. Since nonliteral interpretation is more conventional for the used stimuli, speakers more easily access it in the right context, entirely bypassing the literal reading (*Direct Access Hypothesis*). Now, letting aside the obvious consideration that listing idioms as long words in the mental lexicon would not properly reflect the flexible VP-like nature that some of them have, another fundamental shortcoming shared by both the Idiom List and the Lexical Representation Hypothesis lies in the requirement of an exact match between an input string and a stored idiom to directly trigger a figurative processing (Cacciari & Glucksberg 1991: 219). If an idiomatic sequence indeed demanded a dedicated processing mode or a complete bypassing of the literal meaning computation, a specific cue would be needed at the beginning of the string to activate this type of analysis and directly skip a nonidiomatic reading. Nevertheless, if certain idioms can be, say, passivized or relativized, an input string like *spill the beans* could occur sometimes as it is and sometimes as *the beans have been spilled* or *the beans that have been spilled* and such a specific cue could not be consequently identified. Finally, there are plenty of idioms with begin with the same constituent, despite their different continuations. As Cacciari (2014:

281) reports, Makkai's (1987) dictionary of American idioms lists 132 entries starting with the verb *to take* and 314 idioms starting with the preposition *in.* If idiom processing started with the first word, speakers presented with *take* at the beginning of the sequence would have to simultaneously activate 132 idioms, which is quite improbable. In recent years, Trembley and Baayen (2010) and Trembley and colleagues (2011) have proposed a theory quite similar to Lexical look-up hypotheses for multiword units processing. In their view, a consistent number of frequently employed multiwords are stored and retrieved as wholes in long-term memory without any need to compositional analysis in order to overcome the limitations of working memory (Trembley & Baayen 2010: 3).

A great deal of more recent studies challenge the Lexical look-up vision, preferring a *Non-Lexical* hybrid vision on idiom comprehension (Burt 1992; Cacciari & Tabossi 1988; Cacciari et al. 2007; Fanari, Cacciari & Tabossi 2010; Peterson et al. 2001; Sprenger et al. 2006; Titone & Connine 1994; Holsinger & Kaiser 2013). Cacciari and Tabossi (1988) run a cross-modal priming experiment in which subjects hear sentences ending with either a literal or an idiomatic expression, differing just by the very last word, e.g. *After the excellent performance, the tennis player was in seventh position* vs. *After the excellent performance, the tennis player was in seventh heaven*. In this case, the idiomatic meaning appears to be available at the end of the string. Repeating the experiment with idioms that are not predictable until the final word (e.g. *The girl decided to tell her boyfriend to go to the devil*), the literal meaning of the idiom final word, in this case *devil*, turns out to be activated immediately, while the idiomatic meaning of the string is found to be accessed 300 msec after the presentation of the stimulus. In a follow-up study, Tabossi and Cacciari (1988) observe that both meanings are immediately activated when the idiom is preceded by a figurative-biasing context. In underlining that access to the literal meaning of the idiom components plays a major role in their processing, these data corroborate previous assumptions about the automatic nature of single words comprehension (Stroop 1935; Miller & Johnson-Laird 1976). Relying on collected evidence, Cacciari and Tabossi (1988) propose the so-called *Configuration Hypothesis*, according to which idiomatic strings are first analyzed word by word until sufficient elements have been collected which suggest that the sequence at hand is indeed an idiom. At this point, the figural interpretation is accessed and applied to the string. The point at which an idiom becomes recognizable, called the *idiom key*, depends on both the previous context (Fanari et al. 2010) and the *predictability* degree displayed by the idiom at hand: for predictable idioms like *be in seventh heaven* it coincides with the first part of the sequence, while for non-predictable

idioms like *break the ice* it is located in the final part of the expression. The idea that individual word meanings are accessed in the process is also compatible with Gibbs and Nayak's (1989) *Idiom Decomposition Hypothesis.* Recalling Nunberg et al.'s (1994) claims about idiom decomposability, they observe that decomposable idioms are understood faster than nondecomposable ones since they are compositionally processed. Additionally, in the offline experiments they perform, they find a correlation between semantic decomposability and acceptability of syntactic modifications, therefore providing psycholinguistic confirmation to Wasow et al.'s (1983) theory:

> "the syntactic behavior of idioms is determined, to a large extent, [by] speakers assumptions about the way in which parts of idioms contribute to their figurative interpretations as a whole" (Gibbs & Nayak 1989: 100)

In addition to this, their results partly tie in with the distinction between abnormally and normally decomposable idioms advanced by Nunberg (1978), leading the authors to affirm that analyzability is not an all-or-nothing matter, but is better depicted on a continuum ranging from fully analayzable to fully opaque expressions. In any case, just like we previously affirmed, such a distinction has been put aside since Nunberg et al.'s (1994) contribution. It must be reported, anyway, that later studies show contrasting evidence on online effects of compositionality (Cutting & Bock 1997; Libben & Titone 2008; Tabossi et al. 2008). As also Maher (2013: 12) notes, Gibbs and Nayak's (1989) theory has received some attention within the realm of formal linguistics by Jackendoff (1997), in contrast to Configuration Hypothesis. Jackendoff (1995) conceives the lexicon as the interface of phonological, syntactic and semantic structure. In his model, idiomatic expressions are represented as entries at the lexical-conceptual level that are associated with phonological, syntactic and semantic information that describe the internal structure of the expression itself (Jackendoff 1997). In the lexical representation of an analyzable idiom like *bury the hatchet*, the syntactic component V is marked with the same index *x* of the semantic component RECONCILE, while the syntactic component NP is coindexed (*y*) with the semantic chunk DISAGREEMENT (Fig. 8). Such a coindexing is not obviously adopted with non-analyzable idioms (Fig. 9):

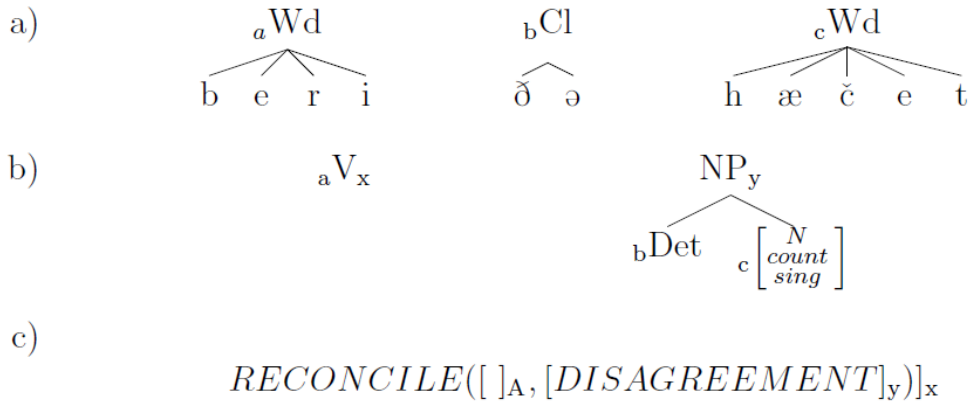



**Figure 8:** representation of the decomposable idiom *bury the hatchet* in the phonological (a), syntactic (b) and semantic (c) structure. (Jackendoff 1997: 168)

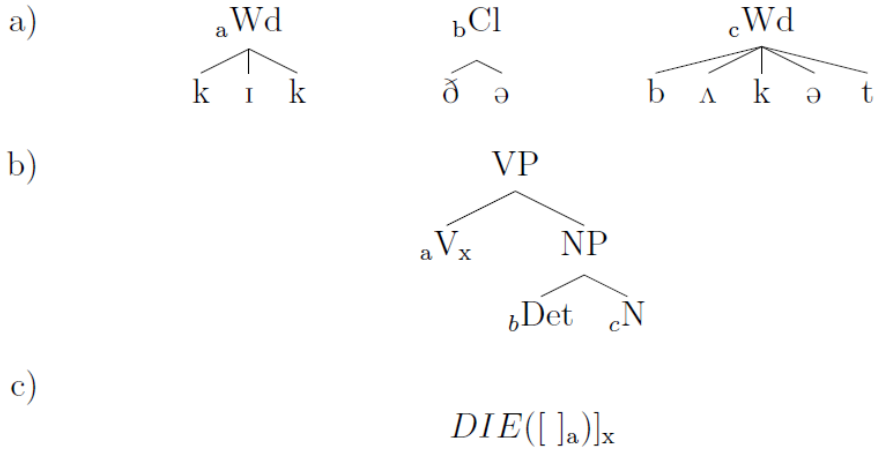



**Figure 9:** representation of the nondecomposable idiom *kick the bucket* in the phonological (a), syntactic (b) and semantic (c) structure. (Jackendoff 1997: 168)

A predictable drawback of this kind of representation is that it does not predict the different degrees of transformability that each analyzable idiom has, since it just provides for differentiating them from nonanalyzable ones with the coindexing device, without further specification (Maher 2013: 13). The distinction between abnormally and normally decomposable idioms (Nunberg 1978) is neglected as well.

Among the existing theories on idiom production, the *Superlemma Hypothesis* (Cutting & Bock 1997; Sprenger et al. 2006) shares the same nonlexical vision of the Configuration and Idiom Decomposition Hypotheses, in spite of the inevitably different perspective taken by production and comprehension theories. While during comprehension the hearer must select between two competing meanings, i.e. idiomatic and literal, on the basis of the context and other factors, in the production process such an ambiguity is not present, since

the speaker clearly knows the message he/she wants to convey (Sprenger et al. 2006). According to Cutting and Bock (1997), who explain idiom production within Dell's (1986) and Levelt's (1989) linguistic production models, this message is represented by a unitary entry at the level of lexical concepts, since idioms, although composed of single words, have a distinct and unitary meaning that cannot even be exactly paraphrased sometimes. Crucially, semantic composition is addressed at the interface with the conceptual and not the syntactic level: the lexical-conceptual nodes for *kick the bucket* and *meet the maker* are activated by the single concept *die*, while for *pop the question* the node is activated by the two concepts *suddenly* and *to propose*. Once a lexical-conceptual node is activated, this activation spreads bidirectionally, both towards the lexical-syntactic nodes, which correspond to the component lemmas, and towards the corresponding phrase pattern. The entire process is summarized in the following schema (Cutting & Bock 1997: 67; Sprenger et al. 2006: 164):

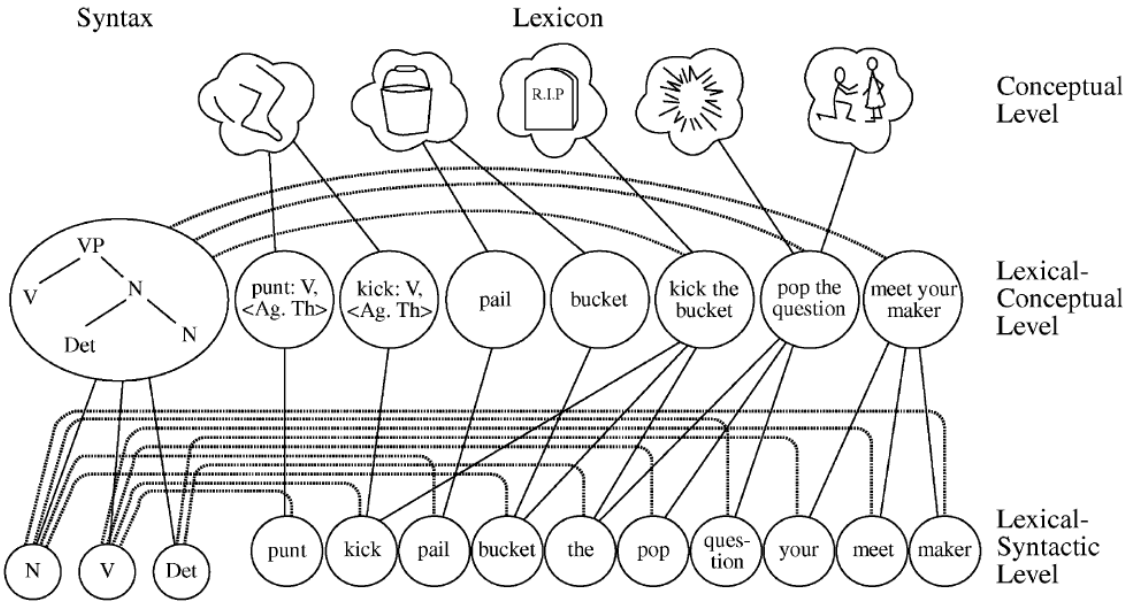



**Figure 10:** Cutting and Bock's (1997:67) lexicon model. All connections are bidirectional (taken from Sprenger et al. 2006: 164)

This account clearly reflects the peculiar status of idioms, which, on the one hand, are associated with a single lexical entry and, on the other hand, make use of the single lemmas stored in the mental lexicon. Depending on the circumstances, a lemma like *breeze*, stored at the lexical-syntactic level, can either be activated by the lexical concept *breeze* to produce a lexical usage of the single word or by the lexical concept *shoot the breeze* to generate the idiomatic expression. It is by this mapping between lexical concepts

and single lemmas that Cutting and Bock's (1997) paradigm provides the Configuration Hypothesis with a productive counterpart. In any case, this theory is elaborated in the light of data elicited via an error induction techniques: participants read two idioms displayed at the same time and produce one of them in response to a cue. This procedure is designed so as to induce phrase blends, which are effectively more frequent when the two simuli are semantically or syntactically equivalent. The same result is obtained with lexical phrases and suggests that idioms in production are both semantically and syntactically analyzed and are not produced as rigid long words. Sprenger et al. (2006) confirm these assumptions with other production experiments showing, for instance, that idioms are produced faster after an identity prime and are completed faster after a prime that is semantically or phonologically related to one of their words. Although these findings tie in with Cutting and Bock's (1997) model in the first place, Sprenger el al. (2006) call into question some aspects of their paradigm, in particular the supposed bidirectionality of the links connecting the processing levels (Dell 1986). If individual lemmas are activated by lexical-conceptual nodes, the top-down relation between them has a semantic nature and states that a given concept, like *hit*, is partly expressed by the corresponding lemma *hit*. When it's an idiomatic lexical-conceptual node to activate its constituent lemmas, this meaning relation holds in the same way: the concept *hit the road* is partly expressed by the lemmas *hit*, *the* and *road*. If this architecture indeed respected Dell's (1986) requisite of bidirectionality, we should detect a revesred semantic conncetion between the lemmas and the upper-level concept as well, but this does not work for idioms: if the lemma *hit* has the meaning of the concept *hit* when it's used literally, the same lemma does not have the meaning of *hit the road* when it's used as part of the idiom. This relation should not be a semantic relation, but instead a part-of relation, which just specifies that *hit* is a chunk of a larger idiom. These issues become evident in the comprehension process, when the above described machinery is accessed in a bottom-up fashion. Sprenger et al. (2006: 176) represent in the following schema the requirement for this double kind of connections between lemmas and lexical-concept, which nonetheless would posit idiom processing and literal phrase processing as two different processes, in contrast with Cutting and Bock's (1997) tenets:

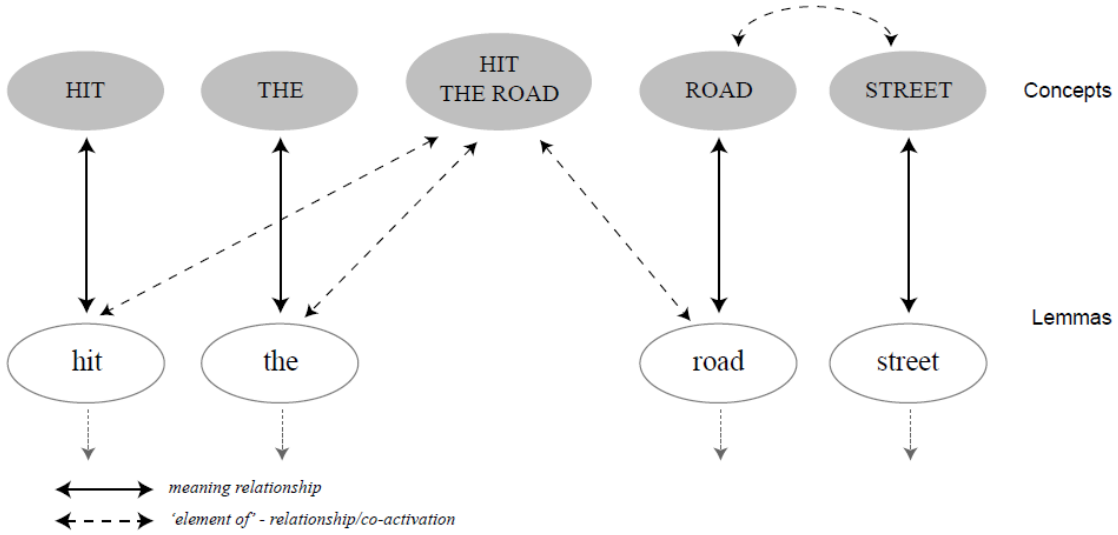



**Figure 11:** *hit the road* represented in Cutting & Bock's (1997) bidirectional model, with two kinds of connections between the lemma and the concept level (Sprenger et al. 2006: 176)

Sprenger et al.'s (2006) solution is to posit a *superlemma* representation of the idiom, which has the function to specify its syntactic structure, at the lemma level, so that part-of connections exist between the single lemmas and the idiom representation within the lemma level, while canonical meaning relations link an idiom to its corresponding lexical concept:

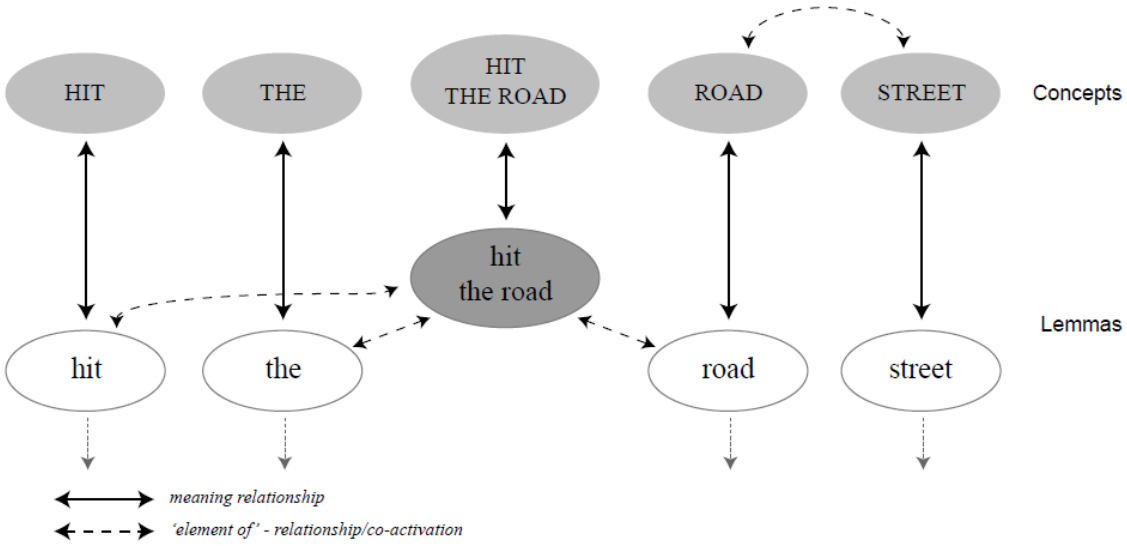



**Figure 12:** representation of *hit the road* in the Superlemma model (Sprenger et al. 2006: 176)

Thanks to the just presented device, lexical-conceptual nodes and lemma nodes are always connected by the same kind of semantic link and the theorized equivalence between idiom and literal phrase processing is consequently preserved. In identifying an idiom with

a lemma, the same lexical competition and lexical selection rules that are at stake in word production come into play. When a superlemma like *hit the road* is activated, its activation spreads to other competing superlemmas and lemmas such as *leave.* Within this process, *hit the road* will be selected only if it exceeds a certain degree of selection probability, equal to the ratio of its degree of activation and the total activation of all lemmas and superlemmas in the system (*Luce's ratio*). Activation then spreads to the individual lemmas, which are in turn chosen via the same competition process. Moreover, since a superlemma works as *"a (phrasal) function over some set of simple lemmas"* (Sprenger et al. 2006: 177) that specifies the syntactic architecture of an idiomatic phrase, it also projects on the individual lemmas the syntactic constraints associated with the idiom. For example, the idiom superlemma *hit the road* deactivates the passive option for the simple lemma *hit* and the construction is therefore not passivizable. The Superlemma Hypothesis bears some resemblance to Katz and Postal's (1963) and Weinreich's (1969) positions on the idiosyncratic nature of idiom syntax, in that all these theories depict idioms with their formal idiosyncrasies stored in their lexical or mental representation and not as being subject to the normal syntactic rules governing the rest of the language. Hence, the only way to properly judge the acceptability of specific syntactic variants for a certain idiom is to have experience of that idiom. Later contributions have nonetheless posed a severe challenge to such assumptions (Konopka & Bock 2009; Tabossi et al. 2009). Tabossi and colleagues (2009) demonstrate that previous exposure to an idiom is not needed to express appropriate acceptability judgments of idiom formal variants and that, as Glucksberg (2001) suggests, the feasibility of certain syntactic transformations is rather conditioned by the more or less appropriate discourse context than a priori stipulated. The final confirmation to the fact that idioms respect the general syntactic principles of language is obtained through the study of bare nouns in passivized idiomatic sentences. Bare nouns are not acceptable in preverbal subject position in Romance languages (Longobardi 2001). If idiom syntax truly obeyed this general rule, passivized idiom phrases with a preverbal bare noun should result ungrammatical to speakers on a par with preverbal bare nouns in literal phrases, notwithstanding their various possible pragmatic contexts. The experimental findings confirm this hypothesis. With a similar viewpoint, Konopka and Bock (2009) perform syntactic priming experiments in which participants are asked to recall sentences they have read in rapid visual representation. Provided that both idiomatic and nonidiomatic phrases appear able to induce stuctural generalizations, sentence formulation is thought to be guided by general sentence creation processes despite the different

semantic compositionality of the items involved.

## *1.4.5. Quantitative approaches to idiomaticity*

### *1.4.6.1. Predicting idiomaticity judgments with corpus statistics: Wulff's (2008; 2009) approach*

Idiomatic expressions have been extensively treated within the sector of corpus linguistics and computational linguistics. The present work springs exactly from this longstanding tradition and aims at devising corpus-driven idiomaticity measure which can predict speaker-elicited ratings regarding a series of psycholinguistically relevant variables, such as idiom predictability, idiom literal plausibility and idiom syntactic flexibility. Wulff's (2008; 2009) contributions represent a predecent for this kind of analysis and their exposition at the beginning of this section reconnects nicely with the survey we just presented on the existing psycholinguistic approaches to idiomaticity. On the other hand, it paves the way for the study that will be described in the following chapters. Here we report a detailed explanation of her study. First and foremost, her work is framed within a constructionist perspective (Langacker 1987; Fillmore et al. 1988; Goldberg 1995; 2006; Croft & Cruse 2004: Bybee 2006; 2010; 2013; Hoffmann & Trousdale 2013), which adopts *Constructions* as units of analysis, namely conventionalized pairings of form and function that are disposed in a *Constructicon* according to their degree of complexity and schematicity (i.e. the opposite of lexical specification). Constructions, in effect, span from simple morphemes to single words, complex words, idioms (both lexically filled and partially lexically filled) and more abstract structural patterns, like the covariational conditional (*the* X-*er the* Y-*er*, e.g. *the cheaper the better*), the ditransitive construction and the passive construction. Here we report a constructicon schema adapted from Goldberg (2006; 2013):

| **Construction** | **Form/Example** | **Function** |
|---|---|---|
| Morpheme | e.g. *anti-*, *pre-*, *-ing* | |

| | | |
|---|---|---|
| Word | e.g. a*vocado, anaconda, and* | |
| Partially filled word | e.g. *pre-N, N-s* (regular plurals) | |
| Complex word | e.g. *daredevil, shoo-in* | |
| Filled idiom | e.g. *take the plunge* | |
| Partially filled idiom | e.g. *stick to X's guns* | |
| Covariational-Conditional | the Xer the Yer, e.g. *the more you eat, the fatter you become* | Meaning: linked independent and dependent variables |
| Ditransitive | Subj V Obj1 Obj2, e.g. *give me a beer; he baked me a cake* | Meaning: transfer (intended or actual) |
| Passive | Subj aux VPpp ($PP_{by}$), e.g. *the boy was hit by a bus* | Discourse function: to make undergoer topical and/or actor non-topical |

**Table 2:** a schema of the *constructicon* (adapted from Goldberg 2006; 2013)

Including in the lexicon also more complex phrasal patterns than just simplex words, that also display different degrees of lexical specification (cf. the distinction between fully specified idioms and idioms with a lexically free slot) and even abstract structural patterns like the passive or the ditransitive construction, this approaches stand in clear opposition to mainstream generative grammar (Chomsky 1957; 1965; 1981) and in particular to its distinction between lexicon and syntax. Their partaking the constructicon is motivated by the fact that also these complex and abstract syntactic templates possess a meaning or at least a function on their own, which is arbitrarily associated with them and must be learnt by the speakers to be properly understood and used. By inserting idiomatic expressions too, these theories also neglect the boundaries drawn by generativism between core phenomena, i.e. those accounted for by canonical syntactic operations, like literal phrases, and peripheral ones, like idiomatic expressions, which are not explicable in light of the basic rules of the syntactic component (cf. Jackendoff & Pinker 2005). In accordance with

psycholinguistic evidence revealing that also fully regular but sufficiently frequent patterns are stored in our mind (Bybee & Hopper 1991; Losiewicz 1992; Bybee 1995; Pinker & Jackendoff 2005), "*even highly transparent expressions that are used sufficiently often to become entrenched in the speaker's mental lexicon qualify as constructions"* (Wulff 2009: 133; cf. Goldberg 2006: 64). Noteworthily, the fact that every member of the constructicon possesses an idiosyncratic semantics or function could lead us to think, and with good reason, that the very notion of idiomaticity is not restricted to idioms in the canonical sense, but is *de facto* applicable to every construction. However, as Wulff (2009: 133) remarks, idiomaticity effects can be best detected and studied on idioms probably due to their reduced schematicy and complexity.

The data under investigation are a set of 39 N-VP constructions extracted from the BNC corpus (100 million tokens) like *bear DET fruit, beg DET question, cross DET finger, make DET headway, pave DET way, play DET game* and *write DET letter.* 39 subjects are presented with each of these constructions used in a context and asked to assign them an idiomaticity rating via magnitude estimation (Bard et al. 1996). This method constists in giving to the first construction seen a score that according to the speaker reflects its idiomaticity and to subsequently evaluate the following constructions with reference to this first one. The main strength of magnitude estimation lies in the fact that judgments are not constrained, for instance, within a 1-7 Likert scale, but can be formulated along a whatever wide range. In the instructions sheet, participants are presented with example idioms, are given a working definition of idioms as *"the kind of sentences you typically find in dictionaries or phrase books"* (Wulff 2009: 134) and are requested to mark how much they think each provided expression differs from normal use and should therefore be listed in a dictionary or phrase books. Interestingly, subjects turn out to be consistent in their ratings (Cronbach's alpha = 0.923; Cronbach 1951), not to rely exclusively on construction frequency, with just a moderately high correlation between average normed values and corpus frequency (r Pearson = −0.635), and to assign differences in ratings that tally with theoretical distinctions. Highest ratings are in effect reserved to opaque idioms like *take DET plunge* and *foot DET bill*, middle scores given to metaphors like *see DET point* and *fight DET battle* and lowest ratings assigned to literal expressions like *write DET letter.* At this point, Wulff devises a corpus-based compositionality index insipired by Berry-Rogghe's (1974: 21–22) study on verb-particle constructions. Berry-Rogghe proposes a compositionality index *R* corresponding to the ratio of the number of collocates shared by both the phrasal verb and the particle and the collocates number of the phrasal

verb:

$$\text{(A)}\ R = \frac{\text{no.of collocates of VPC} \cap \text{P}}{\text{no.of collocates of VPC}}$$

This method ties in well with already proposed compositionality measures (Schone & Jurafsky 2001; Bannard et al. 2003) that compare the context of a construction with the context of its parts. Wulff's (2009) refinement of this measure is first of all motivated by the constructionist idea that each consituent word makes its own contribution to the hierarchically superior construction it occurs in (Goldberg 2006: 10). So, for each V-NP expression, the compositionality index is computed for both the verb and the noun and the two values are then summed. In addition to this, each of the two indices is calculated in an improved version, not only quantifying the contribution of a component to the overall meaning of the expression, as in Berry-Rogghe's (1974) *R*, but also considering how much of a component's meaning is reflected in the construction. This second value is called the *share* of a component and represents the number by which the *R* value of a component is normalized before being added up the the other component's index. To understand the meaning of such a measuremen, we can think about the case of *take DET plunge*, wherein *take* shares only a restricted part of its collocates with the idiom, while *plunge* shares almost all its collocates and receives therefore a much higher share value. Each component *W* of a construction *C* is therefore assigned an *extended R-value*, which is equal to the product of its *R-value* (the number of the collocates shared by *C* and *W* divided by the collocates number of *C*) and its *share* value (the number of the collocates shared by *C* and *W* divided by the collocates number of *W*):

$$\text{(B)}\ Extended\ R = R \cdot share = \frac{n\ colls\ C\ in\ colls\ W}{n\ colls\ C} \cdot \frac{n\ colls\ C\ in\ n\ colls\ W}{n\ colls\ W}$$

The results at least partly corroborate theoretical distinctions, in a similar fashion to the elicited judments, in that they detect a compositionality continuum among the dataset, with idioms like *take DET plunge* and *take DET piss* obtaining the lowest values (0.004 and 0.008 respectively), followed by metaphors like *break DET ground* (0.79) and *carry DET weight* (0.137) and finally by literals like *tell DET story* (0.730) and *write DET letter* (0.844).

The second corpus-driven measure developed by Wulff (2009) accounts for the

morphosyntactic flexibility of the V-NP construction. Its rationale, inspired by Barkema (1994), is that we can get a glimpse of the formal behavior of a given string by comparing it with that of the abstract syntactic construction underlying it (in this case the general *V NP* pattern). If a certain V-NP construction appears more formally rigid than V NP constructions in general, this measure would assign it a low score of morphosyntactic flexibility and vice versa. First of all, the morphosyntactic variational dimensions of interest must be singled out. Those chosen by Wulff (2009: 151 ff.) include:

- *syntactic flexibility* (*SF*) regarding the sentence form, which can be declarative active, declarative passive, relative active, relative passive, interrogative active, interrogative passive and so forth; (for the sake of terminology, *SF* defines a *parameter*, while *declarative active, declarative passive* etc. represent the *parameter levels.* The same applies to the following parameters);
- presence of *addition* (*LF_Add*), *attributive adjectives* (*LF_AttrAdj*), *attributive NPs* (*LF_AttrNP*) and *prepositional phrases* (*LF_PP*);
- presence of *relative clauses* (*LF_RelCl*);
- presence and kind of *adverbials* (*LF_KindAdv*), like space adverbials, time adverbials, respect adverbials, contingency adverbials;
- verbal *person* (*MF_Person*), *number* (*MF_NumV*), *tense* (*MF_Tense*), *aspect* (*MF_Aspect*), *mood* (*MF_Mood*) and *voice* (*MF_Voice*);
- NP *number* (*MF_NumNP*) and *definiteness* (*MF_Det*)

A variational profile for each of the 39 target V-NP constructions is then collected from the BNC by measuring how many times it displays a certain parameter level for each parameter. For an expression like *foot DET bill* (109 tokens), for instance, as regards the MF_Tense parameter, the number of times it occurs in the present (45), past (10), future (9) and nonfinite (45) is collected and the same is done for every parameter. The smaller but fully annotated *International English Corpus* (*ICE*) is then used to extract the variational profile of abstract V-NP constructions in general. After that, the behavior of a given expression like *foot DET bill* regarding a specific parameter level is compared with that of general V-NP constructions with respect to the same parameter level. Such a comparison is first carried out by measuring the *observed frequency* of future for *foot DET bill*:

$$\text{(C) } freq_{observed} = \frac{n\ future\ Foot\ DET\ Bill}{n\ Foot\ DET\ Bill}$$

and by comparing it with the *expected frequency* of the same parameter level, that is the frequency with which we would expect *foot DET bill* to occur in the future if it were a typical V-NP construction:

$$\text{(D) } freq_{expected} = n\ Foot\ DET\ Bill \frac{n\ future\ V{-}NP}{n\ V{-}NP}$$

In the above formulas, *n Foot DET Bill* stands for the idiom tokens, *n future Foot DET Bill* for the number of occurrences of the given idiom in the future, *n V-NP* for the total number of baseline V-NP constructions collected in the ICE and *n future V-NP* for the number of occurrences of abstract V-NP constructions in the future tense. The observed and the expected frequency are then subtracted to obtain the variational measure of *foot DET bill* with respect to the *future tense* parameter level. The variational measure for the entire *MF-_Tense* parameter is obtained by repeating the same procedure for every parameter level, namely *past tense, present tense* and *nonfinite tense.* In the case of *foot DET bill*, the obtained variational indices are −16.80 for the parameter level *past*, −20.66 for *present,* 6.69 for *future,* and 30.77 for *nonfinite*. They are squared and then added so that small deviations of the observed frequencies from the expected frequencies contribute much less than bigger deviations to the overall value. Thus we obtain an overall sum of squared deviations (SSD) for tense for *foot DET bill* of 1700.952,with *nonfinite* contributing 946.904 to this overall value, while *future,* which has a restricted deviation of 6.69% from the baseline, contributes just 44.8 to the overall index. The same method is applied to every other parameter, like SF, LV_AdvKind and so forth.

To observe how these 20 corpus-based idiomaticity indices (18 formal flexibility parameters, plus extended R-values and corpus frequency) cluster, a Principal Component Analysis is conducted. The algorithm groups the indices in 8 principal components, among which the first 4, taken together, account for the 55.635% of the variance of the data. Looking at the component loadings of each corpus-based measure on each principal component, Wulff (2009: 143 ff.) observes that component 1, with an eigenvalue of 4.224, comprises the *SF*, *MF_Voice*, *MF_Det*, *MF_Person* and *LF_AttrNP* parameters. Therefore, tree-syntactic flexibility and verbal morphological variability emerge as the variables that

most explain the distribution of the V-NP expressions in the data. Principal component 2 component comprises *MF NumV* and *MF Mood*, the third component *LF Addition* and *LF NoAdv*, while the fourth one encompasses compositionality and corpus frequency. While the prominence given to syntactic flexibility (Gibbs & Gonzales 1985), adverbial modification (Gibbs et al. 1989) and compositionality in describing idiomaticity variation finds confirmation in plenty of theoretical research, the importance of verbal number and mood appears an element of novelty. In any case, it's worth noticing that compositionality, although emerging in one of the four most important components, does not appear to behave as the variable with the highest explanatory power.

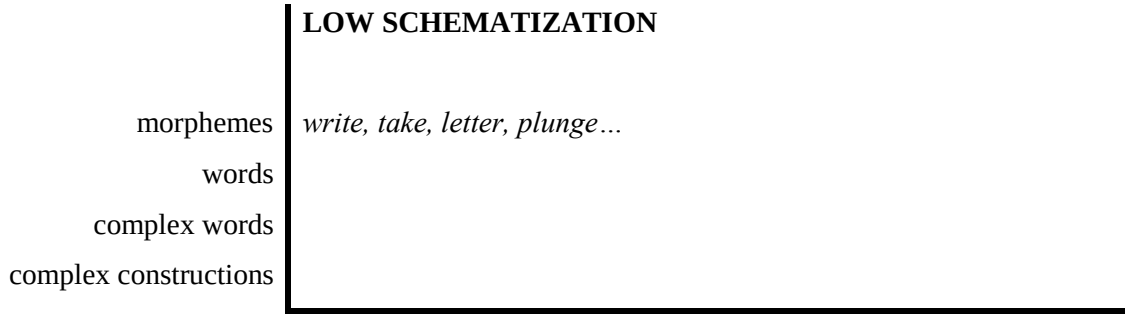
The main research question at the root of Wulff's study is whether the parameters that most explain the variation of the corpus data are also taken into account by subjects to express their idiomaticity judgments. A multiple regression analysis is therefore conducted with the speaker-elicited ratings as a dependent variable and the corpus parameters as predictors. All the predictors, taken together, account for approximately 80% of the variance in the average idiomaticity judgments. Each variation parameter is also assigned a beta weight that quantifies what portion of the variance it covers. Parameters that receive a beta weight higher than 0.22 are considered relevant, since they account for at least 5% of the variance:

| **Variation Parameter** | **Absolute Beta Weight** |
|---|---|
| MF_NumV | 0.757 |
| MF_Mood | 0.695 |
| LF_KindAdv | 0.651 |
| LF_NoAdv | 0.632 |
| Compositionality | 0.578 |
| Tree-syntactic Flexibility | 0.573 |
| MF_Voice | 0.351 |
| MF_Neg | 0.275 |
| LF_Addition | 0.265 |
| Corpus Frequency | 0.209 |

**Table 3:** predictors and respective beta weights in Wulff's (2008; 2009) regression analysis

What comes to the fore is that the parameters that reach a relevant weight also form the most important components in the PCA, with the exception of *MF_Neg*, which forms the

8th principal component, and Corpus Frequency, which remains inferior to the 0.22 threshold. Wulff (2009: 145 ff.) interprets these results in the sense that when a speaker has to evaluate how idiomatic and "far from normal" some phrases are, he/she relies on the same dimensions that best describe the variational behavior of the same phrases in the corpus. Once more, while the importance of tree-syntactic flexibility, lexico-syntactic flexibility (e.g. the occurrence of intervening adverbs in an expression) and compositionality tally with the findings of previous research, the relevance of verbal morphology parameters emerges here for the first time. On top of that, they rank higher than parameters related to the NP slot of idioms and seem indeed to be the most influential factor in human ratings. As concerns compositionality, although it appears among the most relevant parameters, it doesn't receive the most important place in detecting idiomaticity that is usually highlighted in theoretical studies. Wulff (2009: 148) goes on to forestall possible objections about the relevance of verbal morphological parameters being a statistical artifact. If verbal inflection is undoubtedly mandatory and hence not distinctive for idioms, nonetheless we should not exclude that it can be psychologically relevant to perceive idiomaticity. Moreover, if Newman and Rice's (2005) *Inflectional Islands Hypothesis* founded, which predicts that some verbs can be strongly biased towards a particular inflected form, paving the way for grammaticalization, it would corroborate the assumption that speakers are indeed sensitive to the morphological variability of verbs. All in all, merging corpus and experimental evidence shows that idiomaticity is *de facto* a multifactorial and scalar concept and that compositionality and syntactic flexibility are just some of its manifold aspects. How to integrate these findings in the received structure of the constructicon? Wulff (2009: 149) proposes to start from the complex constructions level on the vertical axis, which disposes constructions according to their degree of schematization, and to add a horizontal idiomaticity axis. Expressions located nearer to the vertical axis are more compositional and formally variable, while constructions located on the right are more idiomatic:



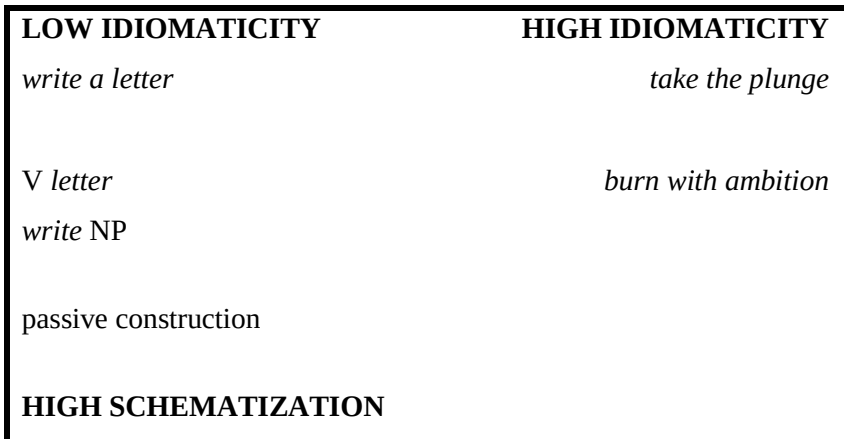



**Figure 13:** incorporation of a horizontal idiomaticity axis in the constructicon by Wulff (2008; 2009)

The horizontal axis is turn composed of a number of layers, each one representing a specific variation parameter, like verbal number, verbal mood, presence of adverbials and the like. These layers are organized in clusters that reflect the principal components previously singled out by PCA. Every complex construction is represented once in each cluster and receives a value for all the parameters included in the cluster. The values obtained for each layer contribute to the overall idiomaticity of the construction. The more idiomatic a construction, the less it is linked with the representation of its component words and vice versa. Since *take DET plunge* is highly idiomatic, it is weakly connected with the representation of *take* and *plunge*. Conversely, *write DET letter* has a very low idiomaticity value and is therefore connected not only with the representations of *write* and *letter*, but also with other related items, like *type*, *compose, email* and *paper*. These words could occur in the construction in lieu of *write* and *letter*, making the *write DET letter* a possible subject to future delexicalizations.

#### *1.4.6.2. Computational studies on idiomatic expressions: state of the art*

In what follows we first provide a concise overview of the manifold methodologies and tasks used in the computational research on idiomatic expressions of the past decades. We then go on to present a detailed description of the study conducted by Fazly and colleagues (2009), since it constitutes one of the most exhaustive and insightful contributions on the matter to date.

Previous research on idiomatic expressions has mainly focused on the assessment of

idiom (and generally MWE) fixedness and on idiom detection.

In the first group we find a series of studies exploiting mostly surface features to determine how rigid a certain MWE is in a continuum from fixedness to productivity. Nissim and Zaninello (2013) compare inflected and lemmatised forms of complex nominals to estimate their rigidity, taking into account the proportion of elements that undergo variation in a given MWE. Zeldes (2013) draws on Baayen's (1992) treatment of morphological productivity and assesses the fixedness of a verb syntactic slot using the number of its hapax noun fillers. Squillante (2014) resorts to frequency-based measures of interruptibility, inflection and lexical substitutability to distinguish multiword units (e.g. *testamento biologico* 'living will', *valor militare*, 'military valour') from lexical collocations (e.g. *carriera solista* 'solo career', *sito ufficiale* 'official website'). Another body of work has dealt with idiom detection in terms of *type* and *token classification*.

Type classification consists in separating potentially idiomatic constructions (e.g. *spill the beans*) from constructions that can only have a literal meaning (e.g. *write a letter*). Such a task can be carried out just relying on superficial features like metalinguistic markers (e.g. proverbially, literally) and quotation marks (Gralinsky 2012). Nonetheless, most research has focused on those linguistic properties that typically distinguish idioms from literals, namely compositionality, lexical fixedness and morphosyntactic and syntactic fixedness. Tapanainen et al. (1998) compare the frequency of a target noun as object with the number of verbs that appear with that object, assuming that objects of idiomatic constructions occur with just one or a few verbs at most. Mc-Carthy et al. (2003) focus on verb-particles constructions, finding a strong correlation between human compositionality judgments and thesaurus-based measures of the overlap between the neighbors of a phrasal verb and those of its simplex verb. Other studies exploit collocational measures (Smadja 1993) or distributional methods that determine the similarity between a given WoC and its components (Baldwin et al., 2003). Venkatapathy and Joshi (2005) combine collocational and distributional measures by means of a SVMbased ranking function that ranks V-N combinations according to their compositionality. Fazly and Stevenson (2008) include in their distributional analysis also the distance between the idiom as a whole and a verb that is morphologically related to its noun constituent, e.g. between make a decision and decide. Muzny and Zettlemoyer (2013) propose a supervised technique for identifying idioms among the Wiktionary lexical entries with lexical and graph-based features extracted from Wiktionary and WordNet. Lin (1999) classifies a phrase as non-compositional if the mutual information between its

components is significantly different from the mutual information of its variants. Each of these alternative forms is obtained by replacing one word in the original phrase with a semantic neighbour. Evert et al. (2004) and Ritz and Heid (2006) use frequency information to determine the preferred morphosyntactic features of idiomatic expressions, while Widdows and Dorow (2005) resort to Hearst (1992)'s concept of lexicosyntactic pattern and extract asymmetric combinations such as A and/or B which never occur in the reversed order B and/or A in their corpus. Such a fixed linear order emerges as a clue of various kinds of relationships between the lexemes pairs, among which idiomatic ones. Bannard (2007) studies syntactic variability of VP idioms, in the form of determiner variability, internal modification via adjectives and passivization. Conditional PMI is used to calculate how the syntactic variation of the pair differs from what would be expected considering the variation of the single lexemes. Fazly et al. (2009) elaborate an overall fixedness measure for V-N combinations that encompasses information on their lexical and syntactic flexibility. The former is derived from an improvement of Lin (1999)'s formula; the latter is obtained by comparing the behavior of a given pair to that of a typical V-N pair as regards the definiteness and the number of the noun and the diathesis of the verb.

*Token classification*, on the other side, consists in recognizing whether a certain word combination is used idiomatically or literally in a given context (e.g. *The old man kicked the bucket two years ago* vs. *Entering the junk room, I accidentally kicked a metal bucket*). Although fine-grained differences actually exist among both the idiomatic and the literal usages of an expression, for the purpose at hand idiomatic and literal usages are approximated to two coarse-grained meanings that a given construction can have (Fazly et al. 2009). Idiomatic tokens can be identified in a supervised (Katz and Giesbrecht 2006; Diab and Krishna, 2009; Hashimoto et al., 2006), weakly supervised (Birke and Sarkar 2006) or unsupervised (Fazly et al. 2009; Sporleder and Li 2009; Feldman and Peng 2013) manner. Katz and Giesbrecht (2006) manually annotate all the occurrences of a given German word-combination as idiomatic or literal, compute a meaning vector for each of the two senses and determine which one is closer to the vector of the token they want to disambiguate. Diab and Krishna (2009) insert richer contextual evidence in a token vector: prepositions, determiners and the collocates of the whole paragraph, not just of the sentence, are considered. Hashimoto et al. (2006) manually create a lexicon which encodes lexical and syntactic information useful to identify different kinds of Japanese idioms via string information, knowledge of the different syntactic transformations they may undergo and disambiguation knowledge that singles out those morphosyntactic patterns under

which the construction loses its idiomatic value. In a follow-up study (Hashimoto and Kawahara 2008), they build an idiom corpus as gold standard and use a WSD method that exploits both common WSD features and idiom-related features taken from the previous paper. Birke and Sarkar (2006) make use of a WSD algorithm which compares a target sentence containing the verb of interest to two seed sets, one containing non-literal sentences and the other containing literal ones. These feedback sets are built automatically, but their sentences contain synonyms of the target expressions extracted from WordNet and the DoKMIE database of idioms and metaphors. Fazly et al. (2009) statistically determine the canonical form(s) for a certain idiom, i.e. its preferred morphosyntactic form, by relying on frequency information and devise an unsupervised classifier that labels a token as idiomatic if it is in the canonical form and as literal otherwise. Sporleder and Li (2009) rely on lexical cohesion to detect idioms in context. Building a cohesion graph in which vertices correspond to the words of a sentence and the edges that connect them are weighted with respect to their semantic relatedness, they observe how the overall connectivity of the graph is affected by the removal of the target expression. In the case of an increased connectivity, the token is classified as idiomatic. In a following improvement (Li & Sporleder 2009) this unsupervised cohesion-based classification is followed by a supervised step that uses SVMs. The features taken into account are the saliency of the context words for the literal interpretation of the token, the semantic relatedness between the token and the contextual words and the connectivity of the cohesion graph. In Li and Sporleder (2010a), a furtherly enriched set of features is proposed among which the words immediately preceding and following the target word, the occurrence of named entities, metalinguistic markers and so on. Li et al. (2010)'s topic model chooses the sense with the highest sense-context probability, where each sense is constituted by a collection of independent words representing its paraphrase and extracted either from idiom dictionaries or via linguistic introspection. Li and Sporleder (2010b) use a Gaussian Mixture Model and assume that two different Gaussians generate literal and nonliteral data respectively. The classification of a tokens is performed by singling out the Gaussian with the higher probability of generating a target instance. Feldman and Peng (2013) apply Principal Component Analysis to extract idioms as semantic outliers and Linear Discriminant Analysis for supervised classification of figurative vs. literal sentences. At the heart of Peng et al. (2014)'s paper lies the same concept of idioms as semantic outliers. In a supervised fashion, they extract topics from paragraphs including VNC used as either idioms or literals via LDA and hypothesize that words appearing as high-ranking

representatives of common topics are less likely to be part of an idiomatic expression in a document. Both Fisher Discriminant Analysis and SVMs with Gaussian Kernels are used as classification schemes. Interestingly, in the light of what Nunberg et al. (1994) claim about the affective value and the nonneutrality of idioms, the emotional arousal associated with each idiom in the dataset is extracted from a database of norms and included in the features set, resulting in a better classification performance. Finally, recent studies on sentiment analysis algorithms have included human-elicited idioms polarity among the features considered, actually obtaining a better performance (Williams et al., 2015).

#### *1.4.6.3. The study of Fazly and colleagues (2009)*

The main goal of this contribution is to attain an unsupervised type-based and token-based identification of idioms by means of statistical measures that best capture their lexical and syntactic fixedness. In a similar fashion to Wulff (2008; 2009), they take into account only expressions composed of a verb and a noun. Firstly, verb-noun pairs occurring at least 10 times are extracted from the BNC corpus. The selected pairs are formed by a set a 28 so-called *basic verbs* that describe fundamental states and activities in the human experience, are highly frequent and polysemous and tend to form idiomatic combinations (e.g. *blow*, *bring*, *catch*, *have*, *give*, *see*, etc.). The *Oxford Dictionary of Current Idiomatic English* and the *Collins COBUILD Idioms Dictionary* are employed as gold standards: collected V-N expressions that are listed in these sources are taken as true idioms also in Fazly and colleagues' dataset. They end up with a development set of 80 idioms and 80 literals and a test set of 100 idioms and 100 literals. To assess the lexical fixedness of an expression, they draw from Lin's (1999) idea. For every verb-noun (*v, n*) pair in the dataset, they extract from Lin's (1998) thesaurus the set of $K_v$ verbs most similar to *v* ($S_{sim}(v) = \{v_i \mid 1 \leq i \leq K_v\}$) and the set of $K_n$ nouns most similar to *n* ($S_{sim}(n) = \{n_j \mid 1 \leq j \leq K_n\}$) and they generate a set of variants by substituting each time the verb or the noun with every word in the $S_{sim}(v)$ or the $S_{sim}(n)$ sets ($S_{sim}(v, n) = \{<v_i, n> \mid 1 \leq i \leq K_v\} \cup \{<v, n_j> \mid 1 \leq j \leq K_n\}$). After parameter settings K is set to 50. PMI (Church et al. 1991) is then calculated for each variant in the set:

$$\text{(E) } PMI(v_r, n_t) = log \frac{P(v_r, n_t)}{P(v_r)P(n_t)} = log \frac{N_{v+n} f(v_r, n_t)}{f(v_r, *) f(*, n_t)}$$

While Lin (1999) labels a *(v, n)* as noncompositional it its PMI is significantly different from those of its variants, Fazly and colleagues assign to every expression a *lexical fixedness* score with the following formula, that confronts the PMI of a certain pair to the average PMI of the variants:

$$\text{(F) } Fixedness_{lex}(v,n) = \frac{PMI(v,n) - \overline{PMI}}{s}$$

The choice of average PMI in the comparison with the variants permits to assign more reliable lexical fixedness scores to idioms (e.g. *take the biscuit*) that have frequently used literal variants (e.g. *make biscuits*). In effect, these would not be labeled as nonliteral according to Lin's (1999) formula, since it requires the PMI of the idiom to be significantly different to the PMI of *all* the variants and *make biscuits* would have a high PMI as well. As for syntactic flexibility, Fazly and colleagues devise a measure that partly recalls Wulff's (2008; 2009) approach and that consists in comparing the syntactic behavior of a given *(v, n)* to that of a typical pair according some relevant dimensions of syntactic variation:

- *passivization*, who occurs for just a set of decomposable idioms and all in all much less frequently than in literal combinations, since it would put focus on an object that is nonetheless nonreferential;
- *determiner type*, since previous literature spots a correlation between determiner flexibility and overall phrase flexibility (Fellbaum 1993);
- *pluralization*, since the referential status of a noun influences its formal flexibility and a nonreferential idiomatic argument is therefore expected to appear just in the singular or the plural form.

Combining these three parameters for all the values they may display, the authors end up with 11 syntactic patterns:

| **Passivization** | **Determiner Variability** | **Number** |
|---|---|---|
| active | Null | singular |
| active | *a*/*an* | singular |

| | | |
|---|---|---|
| active | *the* | singular |
| active | Demonstrative | singular |
| active | possessive | singular |
| active | null | plural |
| active | *the* | plural |
| active | demonstrative | plural |
| active | Possessive | plural |
| active | other | singular / plural |
| passive | Any | singular / plural |

**Table 4:** syntactic patterns considered by Fazly et al. (2009) to compute V-NP syntactic flexibility

The behavior of a typical pair is computed as the prior probability distribution of each pattern *pt*:

$$\text{(G)}\quad P(pt) = \frac{\sum v_i \in V \sum n_j \in N f(v_i, n_j, pt)}{\sum v_i \in V \sum n_j \in N \sum pt_k \in P f(v_i, n_j, pt_k)} = \frac{f(*,*,pt)}{f(*,*,*)}$$

The syntactical behavior of a given *(v, n)* coincides instead with the posterior probability distribution over each pattern *pt* given the *(v, n)* at issue:

$$\text{(H)}\ P(pt|v,n) = \frac{f(v,n,pt)}{\sum pt_k \in P f(v_i, n_j, pt_k)} = \frac{f(v,n,pt)}{f(v,n,*)}$$

The *syntactic fixedness* of a given pair is equal to the Kullback Leibler divergence (Cover & Thomas 1991) between the two syntactic behaviors above:

$$\text{(I)}\ Fixedness_{syn}(v,n) = D(P(pt|v,n)||P(pt))$$
$$= \sum_{pt_k \in P} P(pt_k|v,n) log \frac{P(pt_k|v,n)}{P(pt_k)}$$

Lexical and syntactic fixedness indices are then summarized in an overall fixedness measure obtained via their weighted combination:

$$\text{(L)}\ Fixedness_{overall}(v,n) = \alpha Fixedness_{syn}(v,n) + (1-\alpha) Fixedness_{lex}(v,n)$$

After parameter setting, α is set to 0.6. To evaluate their measures, Fazly and colleagues calculate each of the 3 indices for each verb-noun pair in the dataset and order the pairs according to each of the received scores. Constructions located above the median are assumed to have been labeled as idiomatic, those below as literals. In the following table, Accuracy and Relative Error Rate Reduction are used to evaluate the classification performance, namely the goodness of the indices in telling apart potentially idiomatic and only literal constructions, whereas Interpolated Average Precision at 20%,50% and 80% is used to evaluate the retrieval performance, that is the goodness of the system to rank idiomatic pairs before the literal ones[3]. Lexical, syntactic and overall fixedness indices are evaluated against PMI and Smadja's (1993) method, which measures the fixedness of a pair by quantifying how much the relative position of the component words varies across their occurrences together.

| **Measure** | **%Accuracy** | **%ERR** | **%IAP** |
|---|---|---|---|
| PMI | 63 | 26 | 63.5 |
| Smadja | 54 | 8 | 57.2 |
| $\text{Fixedness}_{\text{lex}}$ | 68 | 36 | 75.3 |
| $\text{Fixedness}_{\text{syn}}$ | 71 | 42 | 75.9 |
| $\text{Fixedness}_{\text{overall}}$ | 74 | 48 | 84.7 |

**Table 5:** evaluation of the classification and retrieval performance of the flexibility measures proposed by Fazly et al. (2009)

Generally speaking, the three fixedness indices perform better than PMI and Smadja. The difference between PMI and lexical and syntactic fixedness does not approach significance, whereas the difference between Smadja and the two indices is significant ($p < 0.05$). The overall fixedness index obtain scores significantly higher than all the other measures ($p << 0.001$), therefore showing the importance of lexicosyntactic variability measures for capturing idiomaticity over and above collocational (PMI) and positional (Smadja) ones.

[3] Idioms that are classified as idioms are *true positives* (*tp*), literals that are classified as literals are true negatives (*tn*), idioms that are misclassified as literals are *false negatives* (*fn*) and literals that are misclassified as idioms are *false positives* (*fp*). *Accuracy* corresponds to $\frac{tp+tn}{total\ data}$, *Precision* to $\frac{tp}{tp+fp}$ and *Recall* to $\frac{tp}{tp+fn}$. To calculate *Interpolated Average Precision*, for every recall level *r* among 20%, 50% and 80%, the highest precision for every $r' \geq r$ is taken and the three values are then averaged.

Token identification is instead based on the consideration that a certain *(v, n)*, when used idiomatically, is likely to occur in a morphosyntactic fixed template, called *canonical form* (Glucksberg 1993; Riehemann 2001). Given a pair like *(see, stars)*, it results clear that, when used literally, it can occur with whatever number and determiner (e.g. *see a star*, *see stars*, *see some stars*, etc.), while it can occur only at the plural and only without determiner when it is used idiomatically (i.e. *see stars*). The intuition of Fazly and colleagues is that, when a given verb-noun pair is encountered in context in its canonical form, it is likely to be used idiomatically. A method for automatically telling apart idiomatic and literal usage in context of a given expression must therefore find, first and foremost, a procedure for automatically singling out the canonical forms of a given word pair. The canonical forms set for a given *(v, n)* (M) is the set of syntactic pattern which have a z-score (N) superior to the threshold $T_z = 1$.

$$\text{(M) } C(v, n) = \{pt_k \in P \mid z(v, n, pt_k) > T_z\}$$

$$\text{(N) } z(v, n, pt_k) = \frac{f(v,n,pt_k) - \bar{f}}{s}$$

Three methods for token recognition are tested:

- CFORM: it labels a token as idiomatic if it is in the canonical form and as literal otherwise;
- CONTEXT: it combines syntactic and distributional information; given the token of a certain couple *(v, n)*, all the other tokens of the same couple are divided into idiomatic, if they occur in the canonical form, and literal, if they don't occur in the canonical form; after that, the average Jaccard[4] distance is computed between the token and the K nearest idiomatic contexts and between the token and the K nearest literal context; the token is finally labeled as idiomatic or literal according to the nearer context;
- SUP: similar to CONTEXT, but the other tokens of the couple are divided into idiomatic and literals by manual annotation;

To provide a gold standard, 100 tokens are randomly extracted for every verb-noun

[4] Jaccard distance between two vectors *x* and *y* is equal to $\frac{x \cap y}{x \cup y}$ (Manning & Schütze 1999: 299).

construction and the idiomatic or literal meaning of the corresponding sentences is manually annotated. Here we report the accuracy and the rate of error reduction for the three method plus a baseline that always assigns an idiomatic meaning. The results refer to two different portion in the dataset: on the one hand, we have idioms with a high proportion of idiomatic-to-literal uses ($DT_{high}$; 65-90% of idiomatic usages, e.g. *get the sack*), on the other hand we have a group of idioms with a much lower proportion of idiomatic-to-literal uses ($DT_{low}$; 8-58% of idiomatic usages, e.g. *see stars*).

| | | **$DT_{high}$** | | **$DT_{low}$** | |
|---|---|---|---|---|---|
| **Method** | | **%Acc** | **%ERR** | **%Acc** | **%ERR** |
| Baseline | | 81.4 | | 35.0 | |
| Unsupervised | CONTEXT | 80.6 | -4.3 | 44.6 | 14.8 |
| | CFORM | 84.7 | 17.7 | 53.4 | 28.3 |
| Supervised | SUP | 84.4 | 16.1 | 76.8 | 64.3 |

**Table 6:** evaluation of the CONTEXT and CFORM measures for idiom token identification (Fazly et al. 2009)

As regards the baseline, since it always predicts an idiomatic meaning, its performance obviously decreases for expressions that have a low proportion of idiomatic usages. As for the $DT_{high}$ group, the method that relies only on the canonical forms obtains the highest accuracy. The supervised method is slightly lower in the ranking, while the Context method performs even worse than the baseline. This is due to the fact that the procedure for deciding the idiomatic and literal contexts is rather noisy, being based on the canonical forms criterion and not on manual annotation. These contexts of comparison are therefore not literal or idiomatic *stricto sensu* but are just labeled so by the unsupervised algorithm. Moving to the $DT_{low}$ dataset, all the accuracy values decrease, with the only notable exception of the supervised method. The underlying reason is quite evident: since these constructions are rarely used in a figurative sense, their canonical form does not appear very frequently or, when it does, it has anyway a literal meaning and its predictive value for the task at hand is consequently weakened.

To sum up, Fazly et al. (2009) study demonstrates that focusing on lexical and syntactic flexibility to identify idiom types and idiom token is a fruitful criterion, especially with respect to previous methods that primarily concentrate on assessing noncompositionality and associational strength (Smadja 1993; Lin 1999; McCarthy et al. 2003; Venkatapathy &

Joshi 2005). Singling out the canonical forms (Glucksberg 1993; Riehemann 2001) of an expression is a reliable technique for distinguishing whether it is used literally or figuratively in context if it has a high ratio of idiomatic-to-literal usages; on the contrary, the predictive value of canonical forms decreases it the construction is mainly used in a literal sense. Moreover, methods for token classification that combine information on canonical forms and contextual information need to exploit less noisy gold standards for establishing which are the literal and figurative contexts with which a token must be compared. The lexical and syntactic flexibility indices devised in this work, in addition to distributional compositionality measures, are also used by Fazly and Stevenson (2008) to classify literal constructions (*take a gift*), abstract combinations (*take a meaning*), light verb constructions (*take a bow*) and idioms (*take pains*), therefore showing the extensibility of this approach to a wider class of MWEs.

# CHAPTER 2
# WORD COMBINATIONS, P-BASED AND S-BASED METHODS AND SYMPATHY

In the present work, we analyze the formal flexibility and the semantic idiosyncrasy of a sample of 87 Italian idioms extracted from a version of the *La Repubblica* corpus (Baroni et al. 2004), a corpus of Italian newspaper texts composed of about 380 millions of tokens which has been POS-tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency-parsed with *DeSR* (Attardi & Dell'Orletta 2009). The expressions have been extracted in the form of verbal subcategorization frames via *SYMPAThy* (*Syntactically Marked PATterns*) (Lenci et al. 2014; Lenci et al. 2015), a format of data representation that encompasses both surface and syntactic information to derive from a corpus the entire combinatorial space of a certain verbal or nominal lexeme. Both subcategorization frames and idioms can be comprised under the overarching label of *Word Combinations*, a term which refers to all the possible constructions a given lemma can co-occur with. Such combinations can be extracted from corpora both via P-based methods, i.e. methods that focus only on the surface Part-Of-Speech level, and S-based methods, i.e. methods that focus on the more abstract level of syntax.

In this chapter we explore the concept of Word Combinations, especially in light of a *constructionist* and *usage-based* view of the mental lexicon (Langacker 1987; Fillmore et al. 1988; Goldberg 1995; 2006; Croft & Cruse 2004; Bybee 2006; 2010; 2013; Hoffmann & Trousdale 2013), we survey both pros and cons of P-based and S-based methods and conclude with a presentation of SYMPAThy (Lenci et al. 2014; Lenci et al. 2015).

## 2. 1. Word Combinations

Studying the distributional behavior of a word, that is, scrutinizing the different contexts in which it can appear and the various lexemes and more abstract constructions it can combine with, is the only way to gain an exhausting insight of its meaning and function (Harris 1951; 1954; Firth 1957; Lenci 2008; Turney & Pantel 2010). This approach to meaning, clearly epitomized in Firth's (1957: 11) catchphrase *"You shall know a word by the company it keeps"*, dates back to the structuralist post-Bloomfieldian works by Harris (1951; 1954) and has acquired new strength in the last decades thanks to the availability of

increasingly bigger text corpora. Within the field of computational linguistics and cognitive sciences, it has led to the emergence of Distributional Semantics (Landauer & Dumais 1997; Lenci 2008; Turney & Pantel 2010), but we leave the description of this framework to the next chapter. What interests us here, from a more theoretical viewpoint, is the very notion of *combinatory space* of a given word.

Taking an Italian verb like *mettere* "to put", we can first of all observe which are the *subcategorization frames* it occurs in most frequently, like "*mettere* Obj PP_*in*"[5], "*mettere* Obj PP_*a*", "*mettersi* Inf_*a*" or "*mettere* Obj PP_*su*", where "PP_*in/su/a*" indicates a prepositional phrase headed by the prepositions *in* ("in/into"), *su* ("over") or *a* ("at"), "Inf_*a*" signals an infinitive clause introduced by the preposition *a* and *mettersi* is a reflexive form roughly meaning "to start doing something". As for its *selectional preferences*¸ i.e. the semantic classes[6] that are most associated[7] to its argument positions (Resnik 1993; Erk et al. 2010), we notice that its object fillers usually belong to the classes GROUP, ARTIFACT, BODY PART, ATTRIBUTE and NATURAL OBJECT, its PP_in fillers to STATE, ACT, COMMUNICATION and LOCATION and so forth. Interestingly, among the fillers which appear to be most associated with the object slot, we find *mano* "hand", *piede* "foot" and *fine* "end", with which the verb forms the two idioms *mettere mano a* "to lay a hand on", *mettere piede in* "to set foot on" and the light verb construction *mettere fine a* "to put a stop to". Going down the list we also find frequent literal combinations, like *mettere una bomba* "to place a bomb" or *mettere una firma* "to put a signature". What we find out with this quick survey is that the combinatory space of a word is composed by a range of different but interrelated phenomena, like typical subcategorization frames, selectional preferences, multiword expressions and frequent literal combinations. In other words, each lemma is endowed with a specific *combinatory potential* that is defined by a range of *constructions.*

This view is taken by a range of approaches to the Grammar and the Lexicon that have developed since the 80s and have emerged in sharp contrast to the tenets of Mainstream Generative Grammar, namely the *constructionist* approaches (Fillmore et al. 1988;

[5] All these data are taken from *LexIt* (Lenci et al. 2012), a resource for the automatic acquisition and analysis of distributional information about Italian verbs, nouns and adjectives, freely available at the address http://lexit.fileli.unipi.it/. The statistics have been extracted from the *La Repubblica* corpus (Baroni et al. 2004) and the Italian section of *Wikipedia.*

[6] These semantic classes correspond to the top nodes dominating the semantic noun taxonomy in the Italian MultiWordNet (Pianta et al. 2002; cfr. Lenci 2014)

[7] The association measure used in this resource is *Local Mutual Information* (*LMI*) (Evert 2008), a variant of *Pointwise Mutual Information* (*PMI*), that avoids its bias towards overestimating low frequency events. LMI between two events *x* and *y* is computed as follows: $LMI = f(x,y)log_2 \frac{p(x,y)}{p(x)p(y)}$

Goldberg 1995; 2006; Croft 2003; Croft & Cruse 2004; Hoffmann & Trousdale 2013). As we have seen in the first chapter, the main principle underlying Construction Grammar is that Lexicon and Grammar constitute, in fact, a *continuum* of *Constructions*, i.e. conventionalized pairings of form and meaning or form and function. These constructions are organized in a network called *Constructicon* according to their degree of *complexity* and *schematicity*. Let's reconsider the constructicon schema we have proposed in the first chapter:

| **Construction** | **Form/Example** | **Function** |
|---|---|---|
| Morpheme | e.g. *anti-*, *pre-*, *-ing* | |
| Word | e.g. a*vocado*, *anaconda*, *and* | |
| Partially filled word | e.g. *pre-N*, *N-s* (regular plurals) | |
| Complex word | e.g. *daredevil*, *shoo-in* | |
| Filled idiom | e.g. *take the plunge* | |
| Partially filled idiom | e.g. *stick to* X*'s guns*, *bring* X *to light* | |
| Covariational-Conditional | the Xer the Yer, e.g. *the more you eat, the fatter you become* | Meaning: linked independent and dependent variables |
| Ditransitive | Subj V Obj1 Obj2, e.g. *give me a beer; he baked me a cake* | Meaning: transfer (intended or actual) |
| Passive | Subj aux VPpp ($PP_{by}$), e.g. *the boy was hit by a bus* | Discourse function: to make undergoer topical and/or actor non-topical |

**Figure 14:** a schema of the constructicon (adapted from Goldberg 2006; 2013)

As the formal requirement assumed by Goldberg (1995: 4) states:

> C is a construction iff$_{det}$, C is a form-meaning pair <$F_i$, $S_i$> such that some aspect of $F_i$ or some aspect of $S_i$, is not strictly predictable from C's component parts or from other previously established constructions.

constructions basically consist in groupings of words that behave idiosyncratically at some level of linguistic analysis. Some have formal peculiarities, while the majority of them have an unpredictable meaning or pragmatic function (Fillmore et al. 1988; Goldberg 1995). Each member of the constructicon is associated with a set of constraints that specify the lexical, semantic and morphosyntactic characteristics of its slots. These kind of constraints define the degree of *schematicity* and *productivity* of a construction (Bybee 1985; Baayen 1993; Goldberg 1995; 2006; Barðdal 2008; Zeldes 2013; Perek 2015). We have already defined *schematicity* as the opposite of lexical specification. A related concept is *productivity*, that is the number of types that can occur in a construction slot (Goldberg 1995; 2006; Bybee & Eddington 2006; Barðdal 2008; Bybee 2010; 2013; Zeldes 2013; Perek 2014; 2015; to appear). Using a more construction-oriented viewpoint, productivity refers to the number of instances (*constructs*) that can be generated from a certain construction. Patterns like the ditransitive and the passive one display the greatest degree of schematicity, since their slots can be filled by almost any lexical item that matches the required grammatical category (e.g. *give me a beer, she baked me a cake, the store sent me a book* etc. for the ditransitive and *the boy was hit by a bus, the president was killed by a sniper* etc. for the passive). Moving down the schematicity-lexicality continuum, we find constructions like the covariational conditional, which is an example of what Fillmore et al. (1988) call *formal idioms*, and idioms with lexically free slots, in which part of the pattern is lexically rigid, while some slots can be instantiated by different fillers. In the case of the covariational conditional, we could have instantiations like *the faster the better, the older the wiser* or more complex ones like *the more you eat, the fatter you become* and the like; as for partially filled idioms, from *stick to* X*'s guns* we could generate *stick to <u>my</u> guns*, *stick to <u>your</u> guns*, or *stick to <u>his own</u> guns*, while starting from *bring* X *to light*, we could produce *bring <u>the problem</u> to light*, *bring <u>the findings</u> to light*, *bring <u>the facts</u> to light* and so forth. On the other end of the continuum we have constructions characterized by the highest levels of lexicalization, fixedness and

idiomaticity, namely totally filled idioms, compound words, single words and morphemes.

Nevertheless, we need to be cautious while equating the presence of a lexically underspecified slot with unlimited productivity. Goldberg (1995: 79) highlights that the English schema "*drive* NP Adj" appears only with an adjective denotating insanity (e.g. *mad*, *crazy*, *nuts*, *bananas*) and that variants such as **drive someone successful*/*angry*/*happy* are not acceptable to speakers, even though these versions do faithfully preserve the resultative meaning of the construction (*CAUSE-BECOME* <*agent result-goal patient*>). In a computational study, Zeldes (2013) finds out that clusters of English verbs that share a certain degree of synonymy (e.g. *comprehend*, *understand* and *fathom*) display strikingly different levels of lexical variability in their object slots. If differences in register can be at stake in the case just cited, the same discrepancy is shown by register-independent syntactic alternations. *Help* and *start*, for instance, can be completed by both the bare infinitive and the *to*-infinitive; the German *wegen* can manifest either as a preposition or as a postposition; in these cases, one of the two variants show a far richer vocabulary than the other. In the next chapter, when introducing the entropy-based flexibility measures we used in our study, we will review the main statistical measures that have been held as the best predictors of syntactic productivity in the last decades.

In light of these differences in lexical, structural and semantic specification among constructions, we can see the constructicon as vertically organized (Croft & Cruse 2004) according to a *default inheritance network* (Goldberg 1995; 2006; 2013). Notably, this type of inheritance hierarchies have also long been used to describe non-linguistic generalizations (Hudson 1990; Lakoff 1987; Goldberg 2006). To provide an example, the English "P N" pattern, in which a preposition is directly followed by a noun without determiner (e.g. *to/at work*, *in/to prison*, *to bed*), inherits its structure and word order from the more abstract PP construction "P NP", the main difference being that PP construction specifies a NP daughter, while "P N" requires an N daughter (Goldberg 2013). Similarly, argument structure constructions can inherit their structure and word order from the VP construction, the Subject-Predicate construction and the Long-distance Dependency construction. As regards resultative constructions, whose meaning can be represented as "X causes Y to become $Z_{state}$", Goldberg (1995) claims that the intransitive version (43a) of this pattern is linked to the transitive (43b) version via a *subpart inheritance* relationship, since it is a proper subpart of it and does not specify the identity of the agent:

(43) a. *The vase broke.*

b. *Mary broke a vase.*

The transitive version can also be related to the caused motion construction ("X causes Y to move $Z_{path/loc}$") (44b), since the resultative phrase and the path phrase show a similar behavior. They cannot be used in ditransitive sentences (45) and cannot occur together (46) (Goldberg 2013).

(44) a. *He drives me to insanity.*

b. *He drives me to Rome.*

(45) a. **He kicks me the ball unconscious.*

b. **He kicks me the ball against the window.*

(46) **He kicks me unconscious against the window.*

All in all, according to its position in the hierarchical network, each construction can capture a different level of generalization in language. Constructions like the Subject-Object, the Subject-Auxiliary Inversion or the VP one detect broader generalizations than patterns like "*What's* X *doing* Y*?*" (Kay & Fillmore 1999). To the more general constructions it inherits, namely Left Isolation, Subject-Auxiliary Inversion, Subject-Predicate and VP, it adds a specific fixed form and an unpredictable pragmatic meaning of surprise or unexpectedness. Therefore, it counts as a construction, but as a less general one, in that it captures a subregularity in the English grammar (Goldberg 2006: 14).

Many constructionist approaches, although not all of them, include in their tenets a *usage-based* perspective on language (Langacker 1987; Hopper 1987; Barlow & Kemmer 2000; Lieven et al. 2003; Tomasello 2003; Bybee 2006; 2010; 2013; Goldberg 2006; 2013). At the root of this theories lies the assumption that language use creates and, at the same time, shapes its cognitive representation (Perek 2015: 6), therefore rejecting the typical distinction between *competence* and *performance* that we find in Generativism. As speakers encounter utterances in their linguistic use, they categorize them relying on phonetics, semantics and the context of use. New elements are sorted and matched to similar representations that already exist in the mind of the speakers, leading to the emergence of units like syllables, words and constructions. Quoting Bybee's (2006) definition, usage-based models conceive grammar as *"the cognitive organization of one's experience with language".* Although the coinage of the term "usage-based" dates back to

Langacker (1987), this conception of language is already present in the functional-typological approach of Greenberg (1963), Givòn (1979) and many others in the 60s and the 70s, who maintain that grammar emerges through the conventionalization of repeatedly used discourse patterns. What is brought in addition by usage-based linguistics is the recognition of the cognitive processes that permit all of this (Bybee 2013).

As recognized by Goldberg (2013), the adoption of this stances allows Construction Grammar to profitably interface with acquisitional, language processing and language change theories. Instead of assuming that language is made possible by special adaptations of cognitive functions, usage-based views argue that domain-general processes are at stake in it, namely processes that are recognized to function in areas other than language, like vision or neuromotor processing (Elman & Bates 1997; Tomasello 2003; Bybee & Beckner 2009; Bybee 2010; 2013). These include *categorization*, *chunking*, *induction, cross-modal association* and *neuromotor automation.*

Categorization refers to the ability to analyze a certain element as an instance of a more general category (Goldberg 2006): phones, words, semantic and contextual features that are witnessed in linguistic usage are categorized by similarity to existing representations (Langacker 2000; Bybee 2010). Categories composed of tokens of experience that are judged to be similar are called *exemplars* (Pierrehumbert 2001). This concept has been first explored in the domain of phonology: the vowels of *hit*, *swim* and *sip* may be united in a same exemplar, just like the various phonetic manifestations of a single word like *cat*, that can be pronounced differently by different speakers (Pierrehumbert 2001; Bybee 2010). Anyway, there exist exemplars of any kind and any size, spanning from vowels to words, constructions and even entire texts that a speaker may have learnt by rote. They form categories that are organized by similarity, frequency and prototypicality effects (Labov 1978; Nosofsky 1988). In particular, exemplars differ in strength, according to the frequency with which they have been witnessed (*token frequency*): those that are constituted by a large number of tokens are more entrenched, i.e. they are represented more strongly, than exemplars that are seldom experienced and they often form the center of a category. The same reasoning obviously applies to constructions: patterns that have a formal or semantic idiosyncrasy and that are experienced with a sufficient frequency in the input are grouped in the corresponding exemplars. It's important to keep in mind that a construction, by definition, is a pairing of form and meaning and it is by virtue of the domain-general process called *cross-modal association* that speakers are able to associate a phonetic, manual or written form to a semantic meaning. As we have seen, the

constructicon encompasses form-function pairings of various levels of schematicity and productivity. Such a difference can be motivated within a usage-based perspective if we take into account the diverging effects of *type* and *token frequency* (Bybee 1985; Bybee & Thompson 1997; Bybee 2010).

Crucially, linguistic elements at all levels of analysis appear with high degrees of repetitiveness in the input. This constant repetition gives rise to conventionalized categories and to a process of *automation*, whereby frequent sequences are combined in conventional ways leading to fluency in production and perception (Bybee 2002). Grammatical organization is therefore massively influenced by frequency effects (Bybee 2007): the more often a given sequence is encountered, the more entrenched it becomes in the mind of the speaker and the more likely is the speaker to create a unique corresponding cognitive representation, which is accessed directly in the successive uses (*chunking*). The importance of chunking and frequency effects is evidenced by the pervasiveness of fully lexically specified patterns that partake the constructicon. In a usage-based conception, they do not comprise only idioms, but also completely literal groupings such as *What's up?*, *What for?*¸ *Tell me what happened*, *I'm sorry to hear that*, *sooner or later* and the like (Goldberg 2013). Initially, even when a certain string undergoes chunking, it remains analyzable and its component words are still recognizable as words that are present elsewhere in cognitive representation. Langacker (1987) represents analyzability through links from the exemplar of a word within a construction to the general exemplar of the word. Through repeated use, analyzability can be gradually lost and the components of a chunk can break their link with the general exemplars of themselves (Bybee 2003). On the other hand, if we have to account for the existence of constructions with schematic slot, we must assume that some form of *inductive* abstraction does take place from lexically specified exemplars to produce more abstract constructions. Following the *usage-based* acqusitional theory of Tomasello (2003), children form representations of partially underspecified constructions (e.g. *Throw* X) from frequent instances in which a given element remains constant and the other element varies (e.g. *Throw your bottle*, *throw teddy*, *throw the ball*). In quite the same way, the constructicon comprises constructions like "*drive* NP Adj" or "V *the hell out of* N" in which the abstract slots are instantiated by different groups of exemplars, that also guide speakers in extending the construction to new uses. It has been reported, for instance, that the lexemes instantiating the verb slot in the English *way* construction mainly indicate either means (47a) or manner (47b) of motion (Goldberg 1995; Israel 1996):

(47) a. *The mole dug his way out of the tunnel*.

b. *The boxer limped his way out of the ring.*

These verbs form two distinct and well entrenched clusters in the semantic space of the construction, signaling that speakers, throughout the historical development of the *way*-pattern, applied new lexical items to it by relying on item-based analogy with previously attested instances (cfr. also Barðdal 2008). On the flip side, it has been recognized that constructions with higher *type frequency*, i.e. higher number of distinct lemmas, are in general considered more schematic and productive than constructions with lower type frequencies in their free slots (Bybee 1985; Bybee & Thompson 1997), since speakers encounter them in a greater variety of contexts and are less likely to assign specific characteristics to their slots in their inductive abstraction processes.

In usage-based theory, therefore, specific facts about the actual use of linguistic expressions, like frequencies and individual patterns, are stored alongside more abstract generalizations in the mind of the speakers, without low-level items being discarded once schematic generalization have formed (Bybee 2006; 2010; 2013; Goldberg 2006; 2013). Co-occurrence of item-based and general knowledge in cognitive representations has been acknowledged by longstanding research in non-linguistic categorization as well (Posner & Keele 1968; Whittlesea 1987; Medin & Schaffer 1978). Moreover, it constitutes another point of departure from the generative tradition, which holds that linguistic elements are either generated by abstract syntactic rules or stored in a lexicon, which is conceived as a repository for all the irregularities in language (the so-called *rule/list fallacy*; cfr. Langacker 1987: 29).

### *2.1.1. A parenthesis on argument structure constructions*

We have seen that among the various existing kinds of constructions, argument structures play a major role. In the constructionist perspective, they *"provide the basic means of clausal expression in a language"* (Goldberg 1995: 3) and are claimed to exist independently of particular verbs. This is at odds with projectionist theories, that see argument structures as entirely determined by the semantics of the verb (Pinker 1989; Van Valin & LaPolla 1997; Rappaport Hovav & Levin 1998; Levin & Rappaport Hovav 2005).

We know that, from a cognitive viewpoint, verb meanings correspond to complex conceptual structures about recurrent events and situations, about their temporal structure, the actors and other elements at stake and the role that each of them plays in such happenings (Barsalou 1992; McRae, Ferretti & Amyote 1997; Perek 2015). In *Frame Semantics* (Fillmore 1985; Fillmore & Atkins 1992) these conceptual structures are called *semantic frames*, defined as *"some single coherent schematization of experience or knowledge"* (Fillmore 1985: 223). Each word, be it a verb or not, evokes a certain frame and the union of this word with the semantic frame evoked defines a *lexical unit*. Polysemic words have therefore the property to evoke a range of different frames. Similarly, according to the projectionist account of Pinker (1989), the meaning of a verb like *eat* is represented by the following semantic structure:

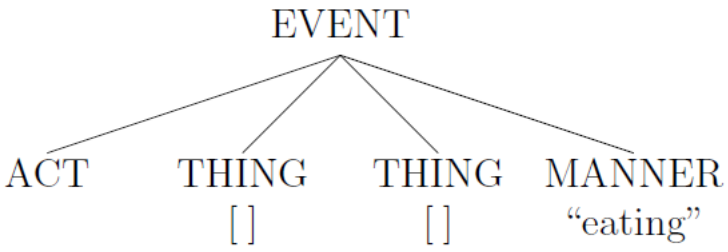



**Figure 15:** semantic structure of the verb *eat* for Pinker (1989: 206)

The verb *eat* denotes an EVENT composed of the ACT predicate and a MANNER component lexically specified by the verb. The ACT predicate, in turn, has two open argument slots, indicated by the THING nodes. These arguments are syntactically realized via linking rules: the first argument is linked to the subject position, while the second argument is linked to the object position:

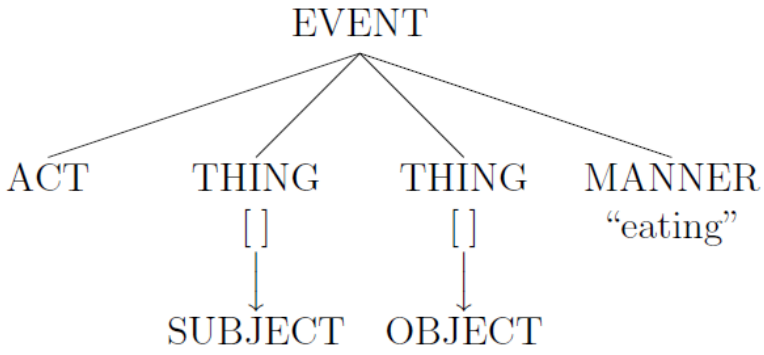



**Figure 16:** argument structure of the verb *eat* following Pinker's (1989) schema

Of course this is just a simple example that catches the basic idea behind projectionist models and linking rules, since a complete explanation of the argument structure characteristics would call for more linking rules acting on more complex semantic structures. Within a similar paradigm, Rappaport Hovav and Levin (1998; Levin & Rappaport Hovav 2005) tackle the issue of *multiple argument realization.* Given that a single verb like *sweep* (Levin & Rappaport Hovav 2005: 188) can appear with different argument structures (48), a strict projectionist approach would have to postulate a distinct verb with a distinct semantic structure for each syntactic context to keep a one-to-one mapping between semantic structures and argument structures:

(48) a. *Terry swept.*
b. *Terry swept the floor.*
c. *Terry swept the leaves into the corner.*
d. *Terry swept the leaves into a pile.*
e. *Terry swept the leaves off the sidewalk.*
f. *Terry swept the floor clean.*

Differently from actual polysemy, though, all these lexical entries for the same verb would denote the same activity, the difference being limited to the outcome of the event. What Rappaport Hovav and Levin (1998) propose is to distinguish the core meaning of a verb (the *constant*), exhibited by the verb throughout all its uses, from the various meanings it shows in context, generated from the interaction of the core meaning with *event templates*. In the case above, the core meaning of *sweep* is defined as a manner constant (<*SWEEP*>), which indicates how the agent acts on the patient in this predicate. This constant can be inserted in event templates that have free "<MANNER>" slots to create an *event structure*. Starting from templates like (49) and (50):

(49) [[ x ACT$_{<MANNER>}$ y ] CAUSE [ BECOME [ y <*STATE*> ]]]
(50) [[ x ACT$_{<MANNER>}$ y ] CAUSE [ BECOME [ z <*PLACE*> ]]]

and inserting <*SWEEP*> in the manner slot, we end up with two event structures realized as a resultative (51) and a caused motion (52) clause, respectively:

(51) *Phil swept the floor clean.*
(52) *Phil swept the crumbs off the table.*

Nevertheless, even this solution falls short of accounting for alternations like (53), wherein both the ditransitive and the prepositional dative pattern denote the transfer of an object to a recipient, without any difference in the event structure:

(53) a. *John gave his mum a present.*
b. *John gave a present to his mum.*

Affirming that *give* has two meanings would also be circular (Goldberg 1995: 10-12), because, on the one hand, we would assume that they have two senses in that they are used in two argument structures and, on the other hand, we would claim that they are used in two argument structures insofar as they display two senses. Another puzzling case is the occurrence of verbs in atypical syntactic patterns, as the intransitive *sneeze* appearing in a caused motion schema (54) or the transitive *light* appearing in a ditransitive pattern (55):

(54) *She sneezed the foam off the cappuccino.*
(55) *Jerry lit us a candle from the emergency kit.*

Assuming that the lexical entries for *sneeze* and *light* also list the meanings "to cause something to move by sneezing" or "to light something with the intention of giving it to someone" seems cumbersome and it is also unlikely that a language of any sort would lexicalize such meanings by means of a separate lexeme (Goldberg 1995: 9-10; Perek 2015: 22-23).

Further evidence comes from the interpretation of nonce verbs: Ahrens (1995) reports that 60% of the participants interpret *moop* in (56) as meaning *give*, with the other subjects assigning a meaning literally or metaphorically related to the concept of transfer (e.g. *tell*):

(56) *She mooped him something.*

In a fundamental contribution, Kaschak and Glenberg (2000) show that speakers demonstrate sensitivity to constructional meaning in interpreting nouns used as verbs in novel ways. (57) is more or less paraphrased as "she used the crutch to pass him the ball",

while (58) is interpreted as “she hit him over the head with a crutch”.

(57) *She crutched him the ball.*
(58) *She crutched him.*

Since *crutch* is not stored in the lexicon as a verb but as a noun, it cannot guide the interpretation of these sentences. This function is committed to the constructional pattern itself, that comes out as having a meaning on its own and as specifying the depicted scene in general. Finally, acquisitional data confirm the implausibility of a lexicalist approach to argument realization. If children actually acquired the argument structure properties of a verb by witnessing the syntactic contexts in which they are used, they would not be able to produce overgeneralizations they have never heard before (Bowerman 1988; Gropen et al. 1989), like the ditransitive use of *say* (59) or the use of *cover* with a direct object theme and a locative PP (60) reported by Bowerman (1988: 79):

(59) *Don’t say me that or you’ll make me cry.*
(60) *I’m gonna cover a screen over me.*

This data clearly signal the existence of general mechanisms that associate argument structures to verbs.

Goldberg (1995) proposes a constructionist analysis of argument structure patterns, regarding them as independent form-meaning pairings that associate an abstract event schema, organized in argument slots, with a morphosyntactic realization. Here we sketch the structure of the ditransitive pattern, both at the semantic and at the syntactic realization level:

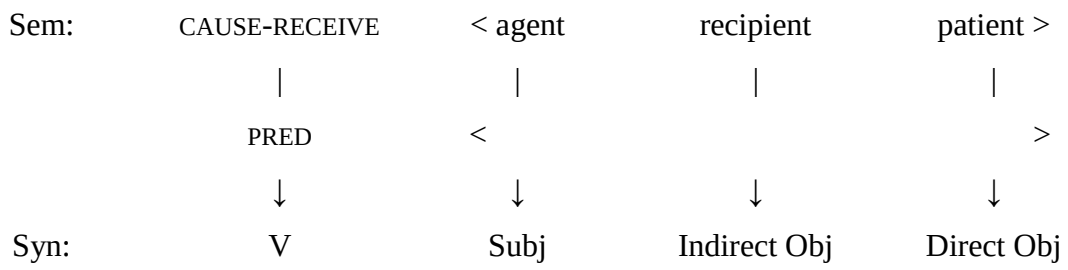



**Figure 17:** schema of the ditransitive construction according to Goldberg (1995: 50)

The *condicio sine qua non* for a verb to partake a construction is *semantic coherence* (Goldberg 1995: 50), which demands that the participant roles defined by the verb meaning are matched with the argument roles of the construction. The prototypical example of such coherence is when a verb meaning includes a constructional meaning, adding further specifications to it. This happens for verbs like *give* and *land* when appearing in the ditransitive pattern. It's worth remarking that the participant roles implied by the verb must be construed as instances of the constructional slots, and don't need to strictly correspond to them. This explains why *tell* can occur in the ditransitive construction as well: it contains a *metaphorical* meaning of transfer, in which the message being told is construed as the given object and the hearer is conceived as a sort of recipient. The account proposed by Goldberg (1995) also permits to motivate the atypical syntactic constructions seen in (54-55). As is known, a caused motion construction implies an agent which causes the motion of a them along the path described by the locative phrase. When an intransitive verb like *sneeze* manifests in this syntactic environment, its single argument is fused with the agent role of the caused motion construction, while the theme and the path are provided by the construction itself:

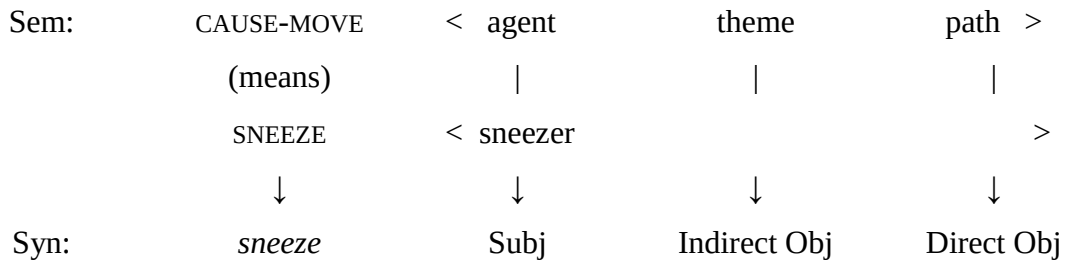



**Figure 18:** fusion of the caused-motion construction with *sneeze* (Goldberg 1995: 54)

## 2.2. P-based and S-based methods for the extraction of Word Combinations

Setting pure theoretical discussions aside, we now turn to the practical problem of how to extract and analyze word combinations by means of computational methodologies.

The combinatory potential of a lexeme can be computationally explored either via methods that exploit surface and pattern-based information (*P-based* methods) or by means

of techniques that capture higher-level syntactic generalizations (*S-based* methods). Their performances varies according to the kind of combinations at hand (Sag et al. 2001; Evert & Krenn 2005).

In both cases, the extraction procedure takes place in a similar way: Part-of-Speech patterns or dependency structures are first of all automatically extracted from a corpus. Secondly, they are ranked according to their frequency and/or various association measures (Evert 2008). This step permits to tell apart noteworthy combinations from sequences of words that are not meaningful (Evert & Krenn 2005; Ramisch et al. 2008; Villavicencio et al. 2007; Lenci et al. 2014; 2015). In the following pages, we cite some relevant works in the field, comparing pros and cons of both methodologies before turning to an exposition of the SYMPAThy framework.

Primarily, POS patterns have been widely exploited for collocation extraction (Smadja 1993; Evert & Krenn 2001; Goldman et al. 2001; Kilgarriff et al. 2004; Krenn &Evert 2005; Wermter & Hahn 2006; Evert 2008). Smadja (1993) proposes a model, called *Xtract*, for the extraction of collocation from general-purpose text, using a combination of n-grams and Mutual Information (Church & Hanks 1989) as a statistical association measure, obtaining a precision of around 80% in identifying collocational units. Evert and Krenn (2001) extract adjective-noun bigrams from a corpus of German law texts and pronoun-noun-verb triples from a portion of the *Frankfurter Rundschau Corpus.* The extracted n-grams are manually annotated as collocational or non-collocational, therefore obtaining a *gold standard* for evaluation, and association measures are then applied to the dataset, resulting in an ordered candidate list for each measure. The employed measures are Mutual Information (Church & Hanks 1989), the log-likelihood ratio test (Dunning 1993), two statistical tests, t-test and $\chi^2$ test, and co-occurrence frequency. The ranking produced by each measure is then evaluated against the annotated set via the so-called *n-best lists* method. It consists in taking the *n* top elements from the ranking produced by each association measure and calculating how many elements in the *n* top ones are labeled as collocations in the gold standard (*precision*) or how many combinations labeled as collocations in the gold standard (*true positives*) are included in the *n* top list (*recall*). Although log-likelihood and t-test emerge as the most reliable association measures in top lists of both 100 and 500 elements, Evert and Krenn (2001) warn that the n-best lists method concentrates only on a small proportion of the entire data and runs the risk to give only partial and misleading results. A possible solution is to plot precision and recall curves for the entire set. Here we report the graphs for preposition-noun-verb data:

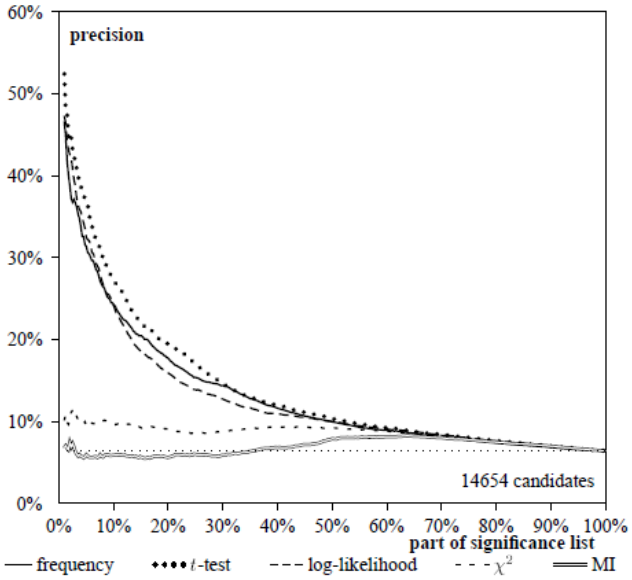



**Figure 19:** precision graphs for the VPN data in Evert and Krenn (2001)

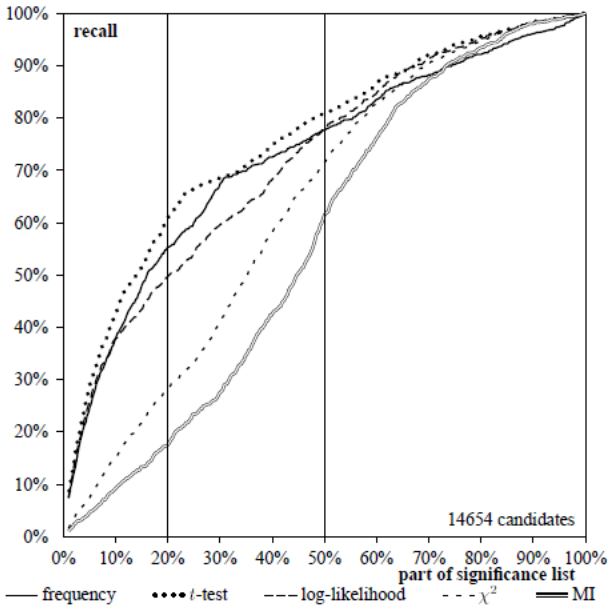



**Figure 20:** recall graphs for the VPN data in Evert and Krenn (2001)

The limited perspective offered by a simple n-best approach is shown by the presence of a number of positions on the x-axis where both precision and recall display almost identical values for every measure. For example, looking at the first 30% of the candidates, we might deduce that t-test and frequency work equally well for extracting preposition-noun-verb collocations. Nonetheless, analyzing the full curves we conclude that t-test is consistently better than frequency. Evert and Krenn (2001) further divide their set

according to frequency strata and observe that the association measures have a different performance in detecting high-frequency and low-frequency collocations. While log-likelihood and t-test neatly emerge as the best performing indices for highly frequent data, the authors report that all the measures have scarce performances in identifying low-frequency collocations, thus bringing into question the widely recognized ability of log-likelihood to handle low-frequency data. On top of that, slight differences are found in the precision and recall graphs between the adjective-noun and the preposition-verb-noun data, since log-likelihood emerges as the best measure in the first dataset, while in the second one it is t-test to provide the best results in collocation identification. In conclusion, their study underlines that association measures applied to POS-specified patterns can be a reliable means to extracting collocations from corpora, although the results are slightly different according to the type of patterns used and the frequency stratum analyzed. A major shortcoming of this method lies, anyway, in the extensive amount of time required by manual annotation of the collocations in the extracted set. In a follow-up contribution, Krenn and Evert (2005) suggest to evaluate association measures against random samples from the full candidate set, so that only a small portion (10 to 20%) of the data is inspected manually.

Superficial patterns extraction has also been extended to Multiword Expressions in general (Ramisch et al. 2008; 2010; Nissim & Zaninello 2013; Nissim et al. 2014; Squillante 2014). Squillante (2014) works on adjective-noun pairs and measures how frequently, with respect to the lemmatized form, each of them occurs in an inflected form, in a form with intervening material in it and in a form with lexical substitution of one of the constituents. Different proportions in inflection, interruptibility and substitutability are used to tell apart literal combinations, multiword expressions and literal collocations in the candidate set. From an online version of the *De Mauro Paravia Dictionary of the Italian Language*, a portion of *GRADIT* (*Grande Dizionario Italiano dell'Uso*) (De Mauro 2000), Nissim and Zaninello (2013) extract and XML-encoded lexicon of MWEs, which is projected on the La Repubblica corpus (Baroni et al. 2004). First of all, a query for the exact quotation forms is attempted. Provided that about 35% of the expressions return zero matches, a flexible search is performed: if a MWE of size *n* is found in the corpus, the pattern $lemma_1$ – $lemma_n$ is searched, exploiting the preexisting corpus annotation for lemmas; otherwise, if the MWE is not found, it is lemmatized and then searched: this leads to the extraction of MWEs like *ghiandola lacrimale* "lacrimal gland" or *faro abbagliante* "high light" that occur only in the plural form in the corpus (*ghiandole lacrimali*; *fari*

*abbaglianti*). Nevertheless, on the one hand, precision decreases with a flexible search, on the other hand, recall is too low for a fixed search. To improve their results, Nissim and Zaninello (2013), in accord with Odijk (2004) and Grégoire (2010), start from the assumption that similar structures, in this case POS patterns, are likely to behave similarly from a morphosyntactic viewpoint. Therefore, concentrating on noun-preposition-noun and noun-prepositional_article-noun patterns, they define *variation patterns* that can be assumed by these sequences while retaining their status of multiword expressions. For instance, starting from the quotation form *casa di cura* "nursing home", labeled as *fix-_fix_fix* since it has every element in his fixed and standard form, we could have a plural inflected form such as *case di cura*, labeled as *flex_fix_fix*, since the first element is flexible with respect to the quotation form. Therefore, recalling the starting assumption, *flex_fix_fix* pattern will be searched also for the other MWEs and so on. What comes to the fore is that variation patterns exhibit great differences in their precision performance. A possible solution mentioned by the authors is to use only those variation patterns that lead to an increase in precision, but this of course disadvantages recall. Since variation patterns are just identified manually and a posteriori in the first part of the experiment, Nissim and Zaninello (2013) also propose a method for the automatic selection of the patterns by comparing the most frequent matches in the corpus for a given MWE to their quotation form, or their respective lemmatized form. These extraction methods are then compared with a range of association measures. While they observe, for instance, that log-likelihood has a worse performance than PMI and Poisson-Sterling measure both for the noun-preposition-noun and noun-prepositional_article-noun data, none of them significantly increases precision with respect to the flexible search method. Nissim et al. (2014) contribution provides additional insights for singling out reliable POS patterns for MWE extraction. Target MWE trigrams containing adjectives are derived from theoretical literature (Voghera 2004; Masini 2012), existing combinatory dictionaries for Italian (Piunno et al. 2013) and their intuition, ending up with the following set:

| **POS-pattern** | **Example** | **Translation** |
|---|---|---|
| ADJ CON ADJ | *Pura e semplice* | Pure and simple |
| PRE ADJ NOUN | *A breve termine* | Short-run |
| PRE NOUN ADJ | *In tempo reale* | (In) real-time |
| ADJ PRE VER | *Duro a morire* | Die-hard |

| | | |
|---|---|---|
| NOUN ADJ ADJ | *Prodotto interno lordo* | Gross national product |
| NOUN NOUN ADJ | *Dipartimento affari sociali* | Social affairs division |
| PRE ADJ VER | *Per quieto vivere* | For the sake of quiet and peace |
| VER PRE ADJ | *Dare per scontato* | To take for granted |

**Table 7:** POS-patterns set identified by Nissim et al. (2014) on the basis of theoretical literature, combinatory dictionaries and personal intuitions

Trigrams containing adjectives are then derived from the La Repubblica corpus (Baroni et al. 2004) and ranked according to raw frequency, log-likelihood and Poisson-Stirling measure. The obtained lists are compared to the pre-identified set, showing that theory-driven and corpus-driven extraction can be profitably combined, providing complementary suggestions. First of all, all the pre-identified configurations are retrieved by the association measures. Secondly, additional configurations are favored in the corpus-driven ranking, like ADJ PRE NOUN (*ospite d'onore* "special guest"), VER ART ADJ (*essere il solo* "to be the only one") or NOUN PRE ADJ (*agente in Borghese* "plain-clothes policeman"). On the flip side, also sequences that are not theoretically sound are suggested by the statistics, either because they represent incomplete sequences, like PRE ART ADJ, or because they are subparts of more complex MWE, like NOUN ADJ PRE (e.g. *concorso esterno in* "external participation in", which lacks the final noun *omicidio* "murder") or ADJ ARTPRE NOUN (e.g. *nazionale del lavoro* "National of Labour", which lacks the initial noun *banca* "Bank").

In wider terms, methods exploiting shallow patterns for the extraction of word combinations demand prearranged specification of the sequences of interest to extract a better candidate set. Nonetheless, as Nissim et al. (2014) show, even after defining a set of reliable patterns, computational methods relying on association measures run the risk to extract combinations that do not correspond *de facto* to MWE, either because they are more restricted than an actual multiword or because they are wider (e.g. *anno di crisi economica* "year of economic crisis" including *crisi economica* "economic crisis"). As a consequence, P-based techniques prove reliable for fixed and short combinations spanning from bigrams to trigrams or 4-grams at most (Lenci et al. 2014). It goes without saying, these approaches are not even suitable for nominal or verbal collocations, light verb constructions or idioms that exhibit considerable syntactic variability (e.g. *pay attention, pay a lot of attention, attention should be paid*, etc.) or are longer and more complex than

simple bigrams or trigrams (e.g. *to run with the hare and hunt with the hounds*).

As concerns syntax-based methods, since the last decade the availability of reliable parsing systems has allowed researchers to exploit syntactic structures for the acquisition of word combinations from corpora (Lin 1998; Blaheta & Johnson 2001; Goldman et al. 2001; Pearce 2002; Korhonen 2002; Schulte im Walde 2008; Erk et al. 2010; Seretan 2011; Lenci et al. 2012; 2014; 2015). Working with French, Goldman et al. (2001) underline the urgency of such an approach, observing that collocational dependencies can sometimes even span along 30 words. Lin (1998) applies information theoretic measures to parse dependency triples, Blaheta and Johnson (2001) resort to log-linear models to capture the associational strength between verb-particle pairs, while Pearce (2001) exploits synonym substitution restrictions to analyze parse dependency pairs. Other works exploit parsed corpora for more abstract and higher level investigations, such as the extraction of subcategorization frames (Korhonen 2002; Schulte im Walde 2008) or the acquisition of selectional preferences (Resnik 1993; Light & Greiff 2002; Erk et al. 2010).

For the Italian language, Lenci and colleagues (Lenci et al. 2012; Lenci 2014) propose *LexIt*, a computational resource for automatically extracting and analyzing distributional information about verbs, nouns and adjectives. Starting from a dependency parsed version of the La Repubblica (Baroni et al., 2004) corpus (ca. 380 millions tokens of newspaper articles) and the Italian section of Wikipedia (ca. 152 millions of tokens), the authors extract the syntactic dependencies governed by a noun, verb or adjective lemma in an automatic fashion. Each of these dependencies constitutes a slot of the target lemma. While there are three kinds of slots that are common to nouns, verbs and adjectives, namely complements (*comp*-), infinitives (*inf*-) and finite clauses (*fin*-), there are also POS-specific slots like subjects (*subj*), direct objects (*obj*), reflexive pronouns (*si*) and predicative complements (*cpred*) for the verbs or the preceding or following modified nouns (*mod-post*, *mod-pre*) for the adjectives. Each target lemma is then associated with the subcategorization frames (*SCFs*) with which it co-occurs. A SCF is constituted by a sequence of syntactic slots headed by a lemma. Consider these examples:

(61) a. *Il governo ha messo la questione in luce.*
"The government has highlighted the issue".
b. *Il governo ha messo in luce la questione.*
"The government has highlighted the issue".
c. *(Il governo) ha messo facilmente in luce la questione.*

"The government has easily highlighted the issue".

All these sentences are assigned to the SCF *subj#obj#comp_in* for the lemma *mettere*, independently from the linear order of the arguments, the presence of adverbial modification and pro-drop. The following expressions:

(62) a. *La legge in vigore sulla sanità.*
"The law in force on healthcare"
b. *La legge in discussione sull'istruzione.*
"The law under discussion on instruction".

are instances of the *comp_in#comp_su* SCF for the noun *legge*, while the SCF *mod_pre#inf_da* for the adjective *difficile* is exemplified by something like (63):

(63) *Una decisione difficile da prendere.*
"A hard decision to make"

The association between a SCF and a lemma is measured in terms of Local Mutual Information. Besides this *syntactic profile*, the target is also paired with a *semantic profile*, which includes the set of fillers that occur in a slot ranked according to their associational strength with the target (the *lexical set*; Hanks & Pustejovsky 2005), and the *selectional preferences*, that is, the semantic classes that are most associated with that slot. All of this is carried out within a completely unsupervised approach and rejecting the traditional distinction between *arguments*, i.e. elements that are needed to complete the meaning of the predicate, and *adjuncts*, i.e. optional elements that are not needed to complete the verbal meaning and that can be erased without the rest of the sentence losing its grammaticality (Dubois et al. 1979; Simone 1990). This distinction proves *de facto* questionable for the defining the meaning of a given lemma and also hard to render into clear-cut, universal criteria.

To sum up, with an S-based resource like this, we can obtain information about the syntactic combinatorics of a given lemma independently from surface phenomena, like word order, occurrence of adjectives and adverbs that separate a verb from its complements and morphosyntactic flexibility. We can find out which are the frames with which it is most associated, which fillers typically occur in a given slot and which

ontological classes are mostly associated with each slot, without the risk of extracting unrelated words (Seretan et al. 2003). Returning to our example lemma *mettere* "to put", we can derive the following overview:

- subj#obj#comp_in
    - OBJ filler: {*piede* "foot", *ordine* "order", *naso* "nose", *meccanismo* "mechanism" …}; {GROUP, ATTRIBUTE, BODY PART, STATE …}
    - COMP_IN filler: {*discussione* "discussion", *scena* "scene", *moto* "motion", *evidenza* "evidence" …}; {STATE, ACT, COMMUNICATION, LOCATION …}

- subj#obj#comp_a
    - OBJ filler: {*fine* "end", *mano* "hand", *rialzo* "rise", *progresso* "progress" …}; {ARTIFACT, KNOWLEDGE, GROUP, EVENT …}
    - COMP_A filler: {*punto* "point", *disposizione* "disposition", *segno* "sign" …}; {LOCATION, ARTIFACT, SHAPE, STATE …}

- subj#si#inf_a
    - INF_A filler: {*piangere* "cry", *fare* "do/make", *ridere* "laugh", *correre* "run" …}

- etc.

Notably, precisely because these approaches neglect superficial patterns, they do not distinguish between word combinations of different kinds that nonetheless display the same syntactic structure: a literal sequence like *non vedere l'uscita* "not to see the exit" would not be classified separately from an idiomatic one like *non vedere l'ora* "to look forward to". However, if we adopted a P-based approach and we focused on the extraction of negation-verb-determiner-noun patterns from our corpus, we could see via association measures that the latter sequence is ranked higher than the former and hence constitutes a meaningful combination. Another case that would be neglected is a distinction of the sort *gettare acqua su un fuoco* "to throw water on a fire" and the idiomatic *gettare acqua sul fuoco* "to minimize", which is identical to the former with the exception of the definite

article. As we have seen, an S-based method would overlook differences on the morphosyntactic axis, categorizing them as two instances of the *gettare#acqua#su_fuoco* SCF.

## 2.3. SYMPAThy: a unified approach to Word Combinations

The idea behind the *SYMPAThy* (*Syntactically Marked PATterns*) (Lenci et al. 2014; 2015) system of distributional knowledge representation is that the insights coming from both P-based and S-based methods must be combined in a unique framework in order to describe all the combinatory properties of the constructicon, without limiting ourselves to a partial view of the phenomenon of word combinatorics.

In our experiments, our point of departure is a version of the La Repubblica corpus (Baroni et al., 2004) POS tagged with the tagger described in Dell'Orletta (2009) and dependency parsed with DeSR, a state-of-the-art stochastic dependency parser (Attardi and Dell'Orletta, 2009). The procedure we are going to illustrate can be applied by using nouns, verbs and adjectives as target lexemes.

First of all, all the occurrences of the target lemma (*TL* henceforth) are extracted from our dependency-parsed corpus. In each sentence the TL occurs in, with respect to any terminal node that depends on the TL, we report in a linear pattern the following information:

- its lemma;
- its POS tag;
- its morphosyntactic features (gender and number for nouns, person, number, tense and mood for verbs);
- its linear distance from the TL;
- the dependency path linking it to TL.

This information is represented in a linear pattern that preserves the linear order of the words in the sentence at hand. Let's start from sentence (64):

(64) a. *I magistrati gettano acqua sul fuoco.*
"The magistrates defuse (scil. the situation)"
Lit. "The magistrates throw water on the fire"

From the dependency parsed sentence in Figure 21:

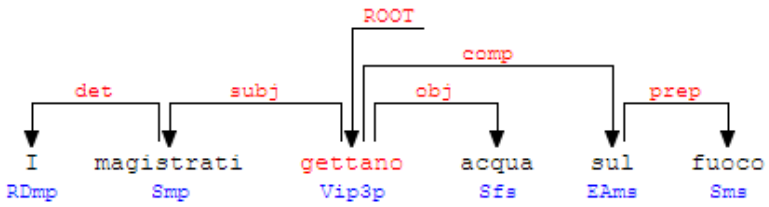



**Figure 21:** dependency parsing of the sentence *I magistrati gettano acqua su fuoco*[8]

we obtain the following patterns using the verb (65) or the noun *acqua* (66) as our TL:

(65) [ $_{\text{TARGET}}$ [ $_{\text{SUBJ}}$ il-r|pm|-2 magistrato-s|pm|-1#H ] **gettare-v|p3ip|0#H**
[$_{\text{OBJ}}$ acqua-s|sf|1#H ] [$_{\text{COMP\_SU}}$ su-ea|sm|2 fuoco-s|sm|3#H ] ]

(66) [ $_{\text{OBJ-1}}$ gettare-v|p3ip|-1#H [$_{\text{TARGET}}$ **acqua-s|sf|0#H** ] ]

Each terminal node is labeled with the pattern *lemma-POS|morphological features|distance from target*. For instance, the pattern *acqua-s|sf|-2* represents an instance of the singular feminine form (*sf*) of the lemma *acqua* "water", that is a noun (*s* as the Italian *sostantivo* "noun") linearly placed one token on the right of the TL *gettare*. As we have said, our patterns captures the syntactic dependencies in the sentence while respecting, at the same time, the linear order of the words in the sentence. Each component of our patterns is delimited by brackets and constitutes an abstraction from the one-to-one dependencies identified by the parser. It can be defined as a meaningful linguistic chunk in which a head member (marked with *H*) is prominent over the others. These non-prominent elements encompass determiners, quantifiers and auxiliaries, namely all the intervening material whose presence or absence is fundamental to assess the degree of formal fixedness of a given expression (see the example *gettare acqua su un fuoco* vs. *gettare acqua sul fuoco* above). In a S-based methods such elements would be overlooked, while in a P-based approach they should be known in advance so as to identify relevant patterns. Moreover, the difference between (65) and (66) highlights how the SYMPAThy format is target-dependent. In each case, the linear order is established with respect to the head

[8] Output provided by the TANL Italian Pipeline: http://tanl.di.unipi.it/it/index.html.

element and also the corresponding syntactic relationships: in (65) *acqua* is linked to *gettare* via an object relationship (OBJ), while in (66) *gettare* is represented in an inverse object relation ($OBJ_{-1}$) with the TL *acqua*. A further element of distinction is the part of the sentence that is analyzed in the pattern: only the constituents that are directly or indirectly governed by the TL and the constituent that governs TL are taken into account, since they are the only relevant elements to characterize the combinatory properties of the TL.

As a further step, abstract syntactic generalizations are derived, namely:

- the syntactic frame the TL occurs in;
- the fillers occupying the slots of the construction.

This results in the following complete pattern, where the TL is associated with the corresponding SCF, with an enriched representation of the SCF including the fillers and finally the same pattern we have seen in (62):

(64) gettare-v subj#obj#comp_su subj:magistrate-s#obj:acqua-s#comp_su:fuoco-s
[ $_{TARGET}$ [ $_{SUBJ}$ il-r|pm|-2 magistrato-s|pm|-1#H ] **gettare-v|p3ip|0#H** [$_{OBJ}$ acqua-s|sf|1#H ] [$_{COMP\_SU}$ su-ea|sm|2 fuoco-s|sm|3#H ] ]

An example of how SYMPAThy works can be offered by the comparison of the patterns containing a transitive construction headed by the verb *gettare* (Lenci et al. 2014). Ranking the object slot fillers by frequency, we find out that the most frequent are *spugna* "sponge", *acqua* "water" and *ombra* "shadow". Widening the focus on entire SCFs, we notice that *gettare#acqua* occurs for the 53.5% of its tokens in the frame *gettare#acqua#su_fuoco*, while in the remaining instances there is a tendency to realize the verb and the object one next to the other, with no morphological variation, with considerable variability in the number, type and fillers of the prepositional phrase. Considering *gettare#ombra*, we observe that the combination is less constrained than *gettare#acqua*. 40% of the tokens instantiate the idiom *gettare un'ombra su* "to cast a shadow on", but, in broader terms, the presence or absence of a determiner, its type, and the occurrence of intervening adverbs and adjectives between the verb and the object is freer. Turning to *gettare#spugna*, we find out that almost all the patterns (98%) are linearly and morphologically fixed, the majority of them being instantiations of the idiom *gettare la spugna* "to throw in the towel".

# CHAPTER 3
# ENTROPIC AND DISTRIBUTIONAL MEASURES OF IDIOM FLEXBILITY

In this third chapter we will be mainly concerned with the computational measures we employed in our study. They comprise entropy-based formal flexibility indices (Shannon 1948; Manning & Schütze 1999; Matthews & Bannard 2010) and distributional semantic indices that capture the semantic idiosyncrasy of our target expressions (Fazly & Stevenson 2008; Mitchell & Lapata 2010; Krčmář et al. 2013). Starting from a review of the major findings achieved in the corpus-driven assessment of idiom flexibility and syntactic productivity, we motivate our choice of *Shannon entropy* (Shannon 1948) as a way to model lexical and morphosyntactic variability and describe the entropic indices used in our experiment. We then describe what *Distributional Semantics* (Lenci 2008; Turney & Pantel 2010) is, how it can be exploited to assess the compositionality of complex expressions in general (Mitchell & Lapata 2010; Baroni 2013) and MWEs in particular (Krčmář et al. 2013) and what distributional measures we implemented in the present study.

## 3.1. Corpus-based assessment of idiom morphosyntactic variability

### *3.1.1. Previous research*

In the first chapter we saw that idioms tolerate various kinds of syntactic modifications, although in an unpredictable fashion (Katz & Postal 1963; Fraser 1970; Wasow et al. 1984; Nunberg et al. 1994). A certain idiom can allow some operations and refuse other ones, but it's extremely difficult, if not impossible, to detect objective criteria whereby this happens. According to Fraser's (1970) hierarchy (*unrestricted variability – reconstitution – extraction – permutation – insertion – adjunction – complete frozenness*), if an idiom tolerates reconstitution (*he laid down the law → his laying down of the law*), it will also permit extraction (*the law was laid down by him*), permutation (*he laid the law down*), insertion (*he laid him down the law*) and adjunction (*his laying down the law*); on the other hand, if it doesn't allow, say, extraction (**a good face was put on by him*), it won't even

undergo reconstitution (**his putting on of a good face*), but its element could be permutated (*he put a good face on*) and so forth. Nevertheless, psychological research (Botelho da Silva & Cutler 1993; Gibbs & Gonzales 1985; Tabossi et al. 2008) has never confirmed the effectiveness of this hierarchy, observing only that adverb insertion is the operation that is most easily accepted (Gibbs & Gonzales 1985; Connine et al. 1992). Nunberg and colleagues (Wasow et al. 1984; Nunberg et al. 1994) maintain that semantic analyzability and syntactic flexibility go hand in hand and such a view is also held by the Idiom Decomposition Hypothesis (Gibbs & Nayak 1989). Anyway, both Cacciari and Glucksberg (1991) and Vietri (2014) have underlined that, actually, totally inflexible idioms do not exist. The former have shown that, under specific discourse circumstances, even idioms that are traditionally associated with complete inflexibility can undergo operations of lexical substitution or insertion that play with their semantics:

(68) A: *Did the old man kick the bucket last night?*
B: *Nah, he barely nudged it.*

(69) A: *By and large, people are well-off these days.*
B: *By and not-so-large! Have you seen the figures on homelessness in America?*

Vietri (2014) analyzes a wide sample of Italian idiomatic expressions using the electronic archives of *La Repubblica* and *Il Corriere della Sera* and the Google query as corpora and finds out that many idiom syntactic variants that are normally considered by the linguistic literature to be unacceptable are *de facto* employed given an appropriate context. For example, *kick the bucket* in English and *tirare le cuoia* in Italian are traditionally classified as nondecomposable idioms, since their meaning "to die" cannot be distributed among their component parts, and according to the aforementioned theories they should be formally frozen. The same should apply to, say, the Italian *sbarcare il lunario* "to make ends meet". Anyway, Vietri (2014) finds many examples in which these expressions are passivized (3) and receive internal modification (4):

(70) a. *Corral fellow thanatologists and let Dead Apple Tours showcase sites where buckets were kicked, dust was bitten, and mortality sponges were squeezed dry.* (Vietri 2014: 8)
b. *Se non sbaglio poi proprio quella sera lì si seppe delle cuoia tirate da Joe Cassano.*

"If I'm not mistaken, that very evening people knew about (lit. the skins pulled by Joe Cassano) the death of Joe Cassano" (Vietri 2014: 40)

c. *Le carriere sono ormai tutte precarie, un giorno dirigenti, un giorno più niente...ed allora una volta sbarcato il lunario, chi se ne frega, aiutalo a cambiare magari...*

"Jobs and careers are all unstable. One day, one is a wealthy manager, the next day one is in a cubicle ... just able to make ends meet (lit. passed the almanac), now why not help that person to change job..." (Vietri 2014: 44)

(71) "*Ragazzi" ha detto "quando Silvio tira le usurate cuoia scendiamo tutti in piazza con le bandiere"?*

"'Guys' he said. 'When Silvio kicks the bucket, (lit. pulls the worn out skins) shall we go out into the streets with the banners'"? (Vietri 2014: 89)

In light of these findings, Vietri (2014) goes so far as to claim the unreasonableness of talking about any sort of exceptionality of idiom syntax: idiomatic expressions appear to be governed by the same syntactic rules that apply to non-idiomatic strings, because even the supposed "ungrammatical" variants appear to be regularly produced by speakers. This consideration reconnects nicely with what Konopka and Bock (2009) observe about the inexistence of an idiosyncratic syntax for idiom in sentence generation. According to Vietri (2014: 49), the main point is not that some syntactic variants are acceptable and other are not, but that some syntactic variants occur much more frequently than other ones. This fact appears to be mainly motivated by register rather than by differences in opacity and decomposability. As for the active-passive alternation, for instance, the passive form is generally speaking much more frequent in formal rather than in informal contexts (Bazzanella 1991; 1994; Biber & al. 1999; Cresti 1999; Biber 2009). As a consequence, an idiom like *tirare le cuoia*, that exhibits more sarcastic, black-humor and therefore informal nuances with respect to the semantically equivalent *salire al cielo* (lit. "to ascend to heaven"), won't be much likely to appear in the passive.

In any case, this notable syntactic versatility must be kept in mind when tackling idioms within a NLP perspective (Sag et al. 2001; Calzolari et al. 2002; Bannard 2007; Wulff 2008; 2009; Fazly et al. 2009). We have seen that to identify idioms within a set of verb-noun combinations in an unsupervised way, Fazly and colleagues (2009) single out 11 relevant patterns of syntactic variation on the basis of verbal diathesis, noun number, and noun definiteness (e.g. $v_{act}$ *det:DEM* $n_{sg}$, $v_{act}$ *det:NULL* $n_{pl}$, etc.). The syntactic fixedness of

a given verb-noun combination corresponds to the difference between the probability distribution over all the 11 patterns for the construction at hand and the probability distribution over the same patterns for a typical verb-noun combination:

$$\text{(O) } Fixedness_{syn}(v,n) = D(P(pt|v,n)||P(pt))$$
$$= \sum_{pt_k \in P} P(pt_k|v,n) log \frac{P(pt_k|v,n)}{P(pt_k)}$$

This is calculated in terms of Kullback-Leibler Divergence (Cover & Thomas 1991). Bannard's (2007) approach to MWEs syntactic flexibility is instead grounded in *Information Theory*, just like the entropic measures we are about to illustrate in this chapter.

This field is introduced and developed by Shannon (1948) in the 1940s. The question that lies at the center of his studies is how to maximize the amount of information than can be transmitted through a noisy communication channel, like a noisy phone line or the like. To answer such a question, Shannon (1948) observe that we must be able to calculate the theoretical maximum for *data compression*, expressed in terms of *entropy* (*H*) and for the *transmission rate*, represented by the *Channel Capacity* (*C*). The probability of errors in the transmission of a message can then be reduced as much as one wants by conveying the information at a slower rate than the Channel Capacity (Manning & Schütze 1999: 60).

Returning to Bannard (2007), the first step in assessing the syntactic versatility of verb-noun pairs is to identify specific kinds of variations that are supposed to occur much less frequently for idiomatic combinations with respect to literal ones. These are:

- variation in the noun definiteness, so that *make waves* becomes *make more waves*, or *strike a chord* becomes *strike chord* respectively;
- adjective, PP or adverb insertion in the NP, so that *break the ice* becomes *break the diplomatic ice*;
- verb passivization, like *call the shots → the shots were called by.*

For each verb-noun combination of interest, the author extracts from a parsed corpus the number of times it occurs with a given variation This is then divided by the token number of that combination to obtain the probability of a given syntactic variation for each verb-

object pair. Noteworthily, to obtain the probability of free variation for a pair, we cannot just compute the product of each of these probabilities, because each phrase has a prior variational probability derived by the variational probability of the component words. Some verbs, for instance, may undergo passivization more likely than others and some nouns can be more prone to adjectival modification than others. The solution is to calculate how much the variational probability for a combination deviates from the variational probability that is expected on the basis of the component words via *Conditional Pointwise Mutual Information* (MacKay 2003).

We have seen that each type of variation is associated with a specific word: passivization is associated with the verb, while adjective insertion and definiteness variability are associate with the noun. Conditional Pointwise Mutual Information expresses the amount of information in bits that a random variable *y* provides about *x* given *z.* For internal modification and definiteness variation, *z* is the object and *y* the verb, while for passivization it is the other way around. Hence, Bannard (2007) measures the information gained about passivization by the addition of the object and the information acquired about internal modification and determiner variation by the addition of the verb with the following formula:

$$\begin{aligned}&\text{(P) } I(x;y|z) = H(x|z) - (H|y;z) = \\ &= -logp(x|z) — logp(x|y,z) \ = \\ &= -logp(x|z) + logp(x|y,z) = \\ &= log\frac{p(x|y,z)}{p(x|z)}\end{aligned}$$

The idea behind it is that if a given variation for a certain pair occurs more likely than we would expect by observing how often that variation occurs with the relevant component word, mutual information will obtain a high score. By adding up the mutual information value for each variation, Bannard (2007) ends up with an overall syntactic variability index for each target pair.

### *3.1.2. Shannon Entropy as a measure of morphosyntactic flexibility*

Wulff (2008; 2009) proposes two kinds of measures for the analysis at hand. The first

one, which we already illustrated in the first chapter, is based on Barkema (1994). The versatility axes that are investigated comprise syntactic flexibility (i.e. whether the sentence is declarative active, declarative passive, interrogative active, etc.), modifiers presence, adverbials presence, verbal morphology and noun number and definiteness. An index for each of these parameters is obtained by subtracting the frequencies of the parameter levels (e.g. *present*, *past* and *future* for the *verbal tense* parameter) for the target V NP construction from the frequency of the same parameter levels for any V NP construction, by squaring the obtained values and by adding them up. A second measure employed by Wulff (2008: 82 ff.) is *Shannon entropy* (Shannon 1948; Manning & Schütze 1999: 61), which computes the average uncertainty of a single random variable.

We define a *random event* as an event that happens unpredictably or that can only be predicted with a certain degree of uncertainty (Manning & Schütze 1999: 40 ff). A *random system* is a system that generates random events. An example of random system could be a 8-sided die roll, for which we could have eight different outcomes, each associated with a probability of 1/8. Each possible outcome is therefore an event that constitutes a subset of the *sample space Ω*, which is in turned composed of all the possible basic outcomes. By calculating the entropy of this variable, we obtain the average length of the binary digits message needed to convey an outcome of this variable. Starting from the general entropy formula:

$$\text{(Q) } H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

we proceed in the following way:

$$\text{(R) } H(X) = -\sum_{i=1}^{8} \frac{1}{8} log \frac{1}{8} = -log \frac{1}{8} = log 8 = 3 \; bits$$

This means that the cleverest and most efficient way to send the result of an 8-sided die roll is to encode it in a binary message composed of 3 digits, like:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 001 | 010 | 011 | 100 | 101 | 110 | 111 | 000 |

All in all, an event with probability *p(i)* is optimally transmitted in a *–log p(i)* bits

message. The negative sign can be moved inside the logarithm, becoming a reciprocal:

$$\text{(S) } H(X) = \sum_{x \in X} p(x) log \frac{1}{p(x)}$$

Please note that the logarithm in the formula has base 2, since the messages are encoded with binary digits: if, on the one hand, given *n* bits, the largest possible number of combinations they can form is $2^n$, on the other hand, given *k* words to encode in bits, $log_2 k$ is the minimum number *n* of bits that are necessary to univocally represent all the words. Notably, entropy has a higher value when all the events have the same probability rather than when one or more events are more likely than others. In the first case, the outcome of the system is more difficult to predict. As a consequence, when a given outcome occurs, it brings with itself a greater amount of information (and, by contrast, a greater reduction of uncertainty) and will need a longer binary string to be encoded. Assuming a second case in which the aforementioned dice is loaded and a given state has a probability of 1/2, while the other seven states have a probability of 1/14 each, the entropy of the system will be inferior to 3 bits. The first event being more likely, the whole random variable is associated with a lesser degree of uncertainty:

$$\text{(S) } H(X) = 7 \frac{1}{14} log 14 + \frac{1}{2} log 2 = 2{,}403677 \; bits$$

Going further, when a single state displays a probability of 1, entropy will be equal to 0, since the occurrence of that state would be certain. It is also worth noting that considering Shannon entropy formula as a sum of *p(x)log(1/p(x))* for every *x* would be wrong. $\sum_{x \in X} p(x)$ is better conceivable as an idiom, that represents an *expectation* and says to compute a weighted average of the rest of the formula, which is a function of *x*. $log \frac{1}{p(X)}$ is therefore weighted by the probability of each *x* (Manning & Schütze 1999: 62):

$$\text{(T) } H(X) = E\left(log \frac{1}{p(X)}\right)$$

Another example proposed by Manning and Schütze (1999:62) is simplified Polynesian. Hawai'ian and the other Polynesian languages exhibit restricted alphabets. We can imagine simplified Polynesian as a random sequence of letters, each of which is characterized by

the following probabilities:

| p | t | k | a | i | u |
|---|---|---|---|---|---|
| 1/8 | 1/4 | 1/8 | 1/4 | 1/8 | 1/8 |

Per-letter entropy results as 2½ bits. The letters above can consequently be encoded as below, with more frequent letters encoded by fewer digits and vice versa:

| p | t | k | a | i | u |
|---|---|---|---|---|---|
| 100 | 00 | 101 | 01 | 110 | 111 |

Once we have explained the basic assumptions lying behind the use of Shannon entropy, we can turn to Wulff's (2008: 82 ff.) proposal. To compute the morphosyntactic fixedness of the target V NP constructions used in her study, Wulff (2008) regards each variational parameter (e.g. aspect) as a system and each parameter level (e.g. progressive, perfective, etc.) as a state of this system. Once more, flexibility (i.e. entropy) is highest when all the parameter levels occur equally often: the more the various kinds of aspect occur with equal probability, the more the V NP construction at hand will be regarded as flexible along this flexibility axis. Conversely, if one aspect prevails, the construction will receive a much lower flexibility score. The following table (Wulff 2008: 84) shows how often the two pairs *draw X line* and *fit X bill* occur in a simple, progressive or perfect tense:

| **Parameter level** | ***n* tokens of *draw X line*** | ***n* tokens *fit X bill*** |
|---|---|---|
| Simple | 194 | 111 |
| Progressive | 30 | 1 |
| Perfective | 86 | 4 |

**Table 8:** frequency distribution of *draw X line* and *fit X bill* for the Aspect parameter

Dividing each cell by the total number of occurrences of the construction of interest (310 for *draw X line* and 116 for *fit X bill*) we obtain the following probabilities:

| **Parameter level** | ***n* tokens of *draw X line*** | ***n* tokens *fit X bill*** |
|---|---|---|
| Simple | 0.626 | 0.957 |

| | | |
|---|---|---|
| Progressive | 0.097 | 0.009 |
| Perfective | 0.277 | 0.034 |

**Table 9:** probability distribution of *draw X line* and *fit X bill* for the Aspect parameter

If we get just even a glimpse of the whole probability distribution, we expect the entropy related to the Aspect parameter for *draw X line* to be higher than the entropy for *fit X bill*, because the three kinds of tenses appear more uniformly in the former, although with a preference for the simple and the perfective tenses. On the other hand, *fit X bill* appears almost always in a simple tense and therefore we would expect a lower entropy. This is actually the case, since $H_{drawXline} = 1.262$ and $H_{fitXbill} = 0.287$. Wulff (2008: 84) further elaborates this entropic measure by computing also *relative entropy*, which consists in the ratio between the observed entropy and the maximum possible entropy of the system. The latter corresponds to the logarithm of the states of the system $|X|$:

$$\text{(U) } H_{rel}(X) = \frac{H(X)}{H_{max}(X)} = \frac{H(X)}{\log_2(|X|)}$$

The rationale behind it is that each variational parameter could be composed of a different number of states, making thus impossible to compare the various entropic values between each other within the same constructions or across different constructions. Relative entropy returns instead a value that spans from 0 to 1 for every flexibility dimension and allows this kind of comparison. Maximum entropy for Table 9 is equal to $log_2(3) = 1.585$ for both the constructions, since they display the same number of states for this parameter. Dividing the observed entropies by the corresponding maximum entropies, we obtain $H_{rel}$(*draw X line*) = 0.797 and $H_{rel}$(*fit X bill*) = 0.181.

While in the case of tree-syntactic and morphological flexibility, it is reasonable to assume that the greater the number of states and the more evenly the tokens are distributed among them, the more flexible the construction is, a different perspective must be assumed while moving to lexico-syntactic flexibility. If we consider adjective insertion as our variational parameter of interest, we should regard as considerably flexible a given expression if the state indicating the presence of an adjective is assigned to most part of the tokens, poorly flexible if the state encoding the absence of an adjective is assigned to most part of the tokens and flexible on average in the intermediate case. The normal entropy formula would contrariwise assign a higher value in the intermediate case and a lower

value in the other two. Wulff (2008: 86) coins a so-called *directional entropy* for the current purpose, that spans from -1 to +1 and assigns negative values if the states without adjectives prevail, positive values when the state encoding the presence of adjectives are the most frequent and a value around zero to the in-between case. To end up with such an index, the normal entropy values are subtracted from 1, so that high entropies result in lower values and vice versa. Then, a positive or negative sign is added according to whether it is the state with intervening material or without intervening material to occur more frequently.

Further insights for the use of Shannon entropy to calculate formal variability come from the syntactic productivity literature. *Syntactic productivity* refers to the number of types a certain construction can occur with (Goldberg 1995; 2006; Bybee & Eddington 2006; Barðdal 2008; Bybee 2010; 2013; Zeldes 2013; Perek 2014; 2015; to appear). To use a more Construction Grammar-related terminology, we could define the syntactic productivity of a construction that possess a certain degree of schematicity as the number of instances (*constructs*) that can be generated from it (Hoffmann & Trousdale 2013). If we dispose the different constructions in the constructicon along a schematicity-productivity gradient, we would start from totally rigid patterns like the idioms *get the sack* and *fit the bill*, pass through partially productive schemas like *bring* NP *to light* or *jog* NP's *memory* and arrive at fully productive patterns like the ditransitive or the passive ones. In her analysis of Italian idioms within the Lexicon-Grammar perspective (Gross 1968; 1984), Vietri (2014) confirms such a difference in productivity by distinguishing cases in which the substitution of a single component results in a loss of the idiomatic meaning (*tagliare la corda* vs. *tagliare la fune*), idioms with a reduced degree of lexical variability (*perdere il treno/tram/autobus*) and idioms with lexically free slots (*voltare le spalle a* NP).

As underlined by Barðdal (2008:29), this conception of productivity as the *extensibility* of a given pattern to new lexemes diverges from the Chomskyan notion of productivity, which is in turn more connected to the concept of linguistic *regularity*. Chomsky (1965; 1981; 1995) focuses on the ability of the speaker/hearer to produce and understand an open-ended set of sentences, including also and especially those never encountered before. In this framework, a system of abstract rules is in charge of generating, through recursive application, this infinite set of syntactic structures. Lexical elements can then freely enter these structures provided that they meet their grammatical requirements. Following contributions, coming mainly from the constructionist and usage-based framework

(Langacker 1987; Goldberg 1995; 2006; Croft & Cruse 2004: Bybee 2006; 2010; 2013; Hoffmann & Trousdale 2013) indeed show that verb argument structures and lexical items do not combine so freely as predicted by the Chomskyan tenet, even when the result would be perfectly acceptable from a semantic viewpoint. Goldberg (1995: 79) observes that in the "*drive* NP Adj" construction, the adjectival slot can only be instantiated by lemmas denoting insanity (e.g. *crazy*, *mad*, *nuts*) and that expressions like **drive someone angry/happy/sick* sound unacceptable to speakers, even though the resultative meaning of the construction *CAUSE-BECOME <agent result-goal patient>* would be theoretically preserved. Zeldes (2013) notices how groups of English verbs that are characterized by a more or less elevated degree of synonymy (e.g. *comprehend*, *understand* and *fathom*) considerably differ in the number of different fillers that can occupy the object slot even when their token frequency is the same. If in this case the difference could be explained in terms of register difference, the same explanation would seem inadequate for verbs like *help* and *start* that can occur both with the *to* infinitive and with the bare infinitive notwithstanding the register and that anyway display a preference for one of the two realizations.

First of all, Goldberg (1995) and Bybee and Thompson (1997) maintain that syntactic productivity depends on the *type frequency* of a construction, that is, the number of lexemes that occur in its lexically free slots. In other words, speakers would be more prone to employ a construction with new lexical items if they have already witnessed it being used with a significant number of lexemes rather than with a restricted set of them. Following Goldberg (2006: 99), Barðdal (2008: 34 ff.) takes also the *semantic variability* of a construction into account and claims that syntactic productivity is a function of the type frequency of a construction, its semantic coherence and an inverse correlation between the two. Productivity emerges thus as a gradient phenomenon. On one side of the continuum, we find constructions that occur with a wide range of semantically different lexemes and that will be liable to new instantiations only if they are witnessed with a significant number of types (*schema-based productivity*). Moving along the continuum we find constructions with a lower type frequency and that are productive only if these items are semantically close, arriving at the other end, where we find the rare cases of analogical extension, whereby patterns that are instantiated by a single type are extended to another one, resulting in a new construction semantically equivalent to the first one (*analogy-based productivity*). Analyzing a sample of 107 recent Icelandic verbs related to the field of Information Technology, Barðdal (2008: 78 ff.) notices that 64% of them is attracted by

the Nominative-Accusative construction, that has the highest type frequency and covers the greatest number of semantic classes in Icelandic, while only 36% of them take the Nominative-Dative construction, which displays both lower type frequency and semantic variability. An example of analogy-based extensions is the English verb *dawn*, the enters Icelandic as the prepositional verb *dona uppi* "to be forgotten" in analogy with the Icelandic *daga uppi*, that has the same meaning and is composed of the word *daga* "dawn". Noteworthily, Barðdal's (2008) conception of schema-based and analogy-based productivity as two sides of the same coin is at odds with Tomasello's (2003) acquisitional usage-based theory. According to the latter, schematization leads a child to abstract partially underspecified constructions (e.g. *Throw X*) from a series of instances in which an element remains fixed and the other changes (e.g. *Throw the ball, throw teddy, throw your bottle*), while analogy leads to the abstraction of entirely schematic patterns (e.g. the ditransitive patterns) on the basis of the observation that certain elements in a series of utterances play always the same function (e.g. agent or theme).

Other factors have been demonstrated to be at stake in determining the productivity of a construction, such as the semantic similarity between the novel coinage and the previously attested types (Bybee & Eddington 2006; Suttle & Goldberg 2011), the semantic density of the free slots (Perek 2014; to appear), the pragmatic function of the construction[9] (Perek & Goldberg 2015) and statistical pre-emption[10] (Brooks & Tomasello 1999; Goldberg 1995; 2006; Marcotte 2005; Boyd & Goldberg 2011), but they are not the focus of our current discussion. What deserves our attention, instead, is that in the acquisitional study by Matthews and Bannard (2010), Shannon entropy is proposed as a reliable predictor of syntactic productivity. Acquisitional contributions have extensively tackled the issue of syntactic productivity (Bowerman 1988; Pinker 1989; Tomasello 2003; Bannard & Matthews 2008; Matthews & Bannard 2010; Theakston et al. 2015), aiming at verifying which mechanisms lead the children to extend the linguistic constructions they have learnt to a wider range of lexemes than the ones witnessed in the intrinsically limited parental input. Matthews and Bannard (2010) devise a sentence repetition task, in which twenty-eight 2-year-old and thirty 3-year-old children are asked to repeat 4 word unfamiliar word

[9] In an experiment of artificial language learning, Perek and Goldberg (2015) create two constructions, SOV and Pronoun-SV respectively, that differ in the *givenness* degree of the object and demonstrate that subjects rely on the informative function of the construction when extending their use to new instances, depending on the context in which the new instantiation must be uttered.

[10] Statistical pre-emption is that process whereby, in a context in which a construction A could be sensibly employed (**explain me this*), a learner will conclude that A is not appropriate if a competing construction B is repeatedly and consistently witnessed (e.g. *explain this to me*).

sequences that are identical to familiar word sequences except for the last word. An example could be the repetition of the familiar pattern *out of the* X with an unfamiliar filler, like *out of the pudding*. The unfamiliar sequences used in this experiment are sequences that never appear in a large corpus of child-directed speech, namely the *Max Planck Child Language Corpus* (1.72 million words of maternal speech). The focus of this study is to observe how the statistics of the input influence children's tendency to use some constructions more productively than other ones. The authors start from the hypothesis that children's ability to repeat the target patterns is influenced by their previous exposure to the corresponding constructions. In a previous work (Bannard & Matthews 2008), they observe that children more easily and more correctly repeat the first three words of frequently occurring strings than the first three words of matched and less frequent sequences (e.g. they better repeat *sit in your* when saying *sit in your chair* than when uttering *sit in your truck*). Focusing just on unfamiliar sequences, Matthews and Bannard (2010) notice that the ease of repetition should be affected by the type frequency of the construction of interest if we stuck to the received wisdom on productivity. However, they add that type frequency alone would not take into account the different probability exhibited by each type. Tomasello (2003) maintains that children form the partially underspecified pattern "*throw* X" by witnessing the fixed part *throw* co-occurring with a variety of types, like *ball*, *teddy*, *bottle*, etc. Anyway, in the case a child heard *throw your bottle* 118 times and both *throw the ball* and *throw teddy* only once, he/she would always expect to hear *your bottle* after *throw* and would not identify the fixed part *throw* as productive. By contrast, if the three examples were more evenly distributed, with a frequency of 40 each, the child would be more uncertain about the word following *throw* and would be more likely to identify the pattern as productive. Such a difference can be perfectly captured by measuring Shannon entropy for the X slot, which would be equal to 0.14 in the unequal case and 1.58 in the equal case. In slot entropy, each state *x* represents each filler that can occur in the slot. Subjects are therefore expected to produce more correctly a given string if slot entropy is high and vice versa and the results confirm this hypothesis.

### *3.1.3. Our entropic indices*

Drawing from the studies mentioned so far, we have resorted to Shannon entropy to

calculate the formal flexibility of our target idioms. This measure is applied to the P-based and S-based information extracted from our corpus (see Chapter 2) to obtain a *variational profile* that summarizes:

- the variability of the fillers that instantiate the free slots of those idioms that exhibit them;
- the morphological variability of both the verb and the arguments;
- the variability of the fillers definiteness;
- the variability in the presence of adjectives and PPs modifying the slots, or adverbs modifying the verb;
- the variability in the linear order of the slots with respect to the verb

### LEXICAL ENTROPY

First of all, such variability can emerge at the lexical level, if we consider a partially lexically specified idiom like *gettare luce su X* "to cast light on X" and the different lexical instantiations its free slot may have, like *gettare#luce#su_problema*[11], *gettare#luce#su_vicenda* or *gettare#luce#su_fatto*. Considering all the tokens of *gettare luce su X* in our corpus, how many words, that is *types*, can fill the X-slot and with which probability? Shannon entropy can give a reliable answer to this question. Let's consider the formula once more:

$$\text{(Q) } H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

In this case, the random variable is represented by the lexical variability axis of the construction at hand, while each state *x* is every possible lexical realization of the X-slot in terms of types, say *gettare#luce#su_problema, gettare#luce#su_vicenda* or *gettare#luce#su_fatto* as we said above. *X* represents the set of all the possible states, while

[11] In the following exposition we will represent the idioms in the form of SYMPAthy frames, where each lemma or preposition+lemma is separated by a symbol "#". This representational format is also suitable for underlining that each entropic measure focuses just on a single variational dimension. In the case of lexical entropy, for instance, the alternation between *gettare#luce#su_problema* and *gettare#luce#su_fatto* abstracts away from differences in number and definiteness of the nouns (*gettare luce sul problema* "to cast light on the issue", *gettare luce su un problema* "to cast light on an issue" , *gettare luce sui problemi* "to cast light on the issues", etc.) and considers just the filler variation.

the probability of each one is obtained dividing its raw frequency *f(x)* by the total number of tokens of the underspecified idioms.

Most notably, for every kind of variational dimension, we calculate relative entropy, in agreement with Wulff's (2008) proposal, therefore dividing the observed entropy by the logarithm of the number of the states, in order to make the various entropic measures comparable between each other.

### MORPHOLOGICAL ENTROPY

Another variability axis is the morphological one. Nouns, adjectives, pronouns and articles in a SYMPAThy pattern are associated with information about their gender and number, while verb are associated with information about their mood, tense and person. For each instance of an idiom we can therefore extract the morphological features of its argument slots and then derive some general statistics: out of the 961 tokens of *gettare#acqua#su_fuoco*, for instance, we have *gettare#acqua-FEMININE.SINGULAR#su_fuoco-MASCULINE.SINGULAR* 960 times and *gettare#acqua-FEMININE.SINGULAR#su_fuoco-MASCULINE.PLURAL* just 1 time. This kind of morphological variation is captured by *morphological entropy*, where each state *m* stands for each possible combination of morphological features of all the slots at the same time, like *gettare#FEMININE.SINGULAR#su_MASCULINE.SINGULAR* and *gettare#FEMININE.SINGULAR#su_MASCULINE.PLURAL* for the examples above. It's important to underline that if we are calculating morphological entropy for a partially lexically underspecified construction, each state is just represented by each possible combination of morphological features for all the slots, without considering the different fillers these slots may have. In other words, gettare#ombra-*FEMININE.SINGULAR*#su_vicenda-*FEMININE.SINGULAR* and gettare#ombra-*FEMININE.SINGULAR*#su_questione-*FEMININE.SINGULAR* would count as two instances of the same state *gettare#FEMININE.SINGULAR#su_FEMININE.SINGULAR*. The same applies when we calculate any other kind of entropy, say *order* or *modifiers entropy*, for partially lexically underspecified constructions.

### ORDER ENTROPY

Given a certain construction, we don't find its slots and verbal target always in the same

reciprocal order throughout the corpus. For *gettare#acqua#su_fuoco*, we find *su_fuoco#gettare#acqua* 1 time and *gettare#acqua#su_fuoco* 960 times. Each of these order schemata represent a state of this kind of entropy.

#### DISTANCE ENTROPY

Each element in a SYMPAThy pattern is marked by a number indicating its distance in tokens from the target lemma. *Distance entropy* then captures how variable is the distance from the verbal target of the slots of a construction over all its occurrences in the corpus. Possible states are *gettare#acqua-4#su_fuoco-6, gettare#acqua-1#su_fuoco-4* or *gettare#acqua-1#su_fuoco-3.*

#### ARTICLES ENTROPY

*Articles entropy* captures the variability in the presence or absence of articles determining the slots of a construction and, in the former case, their type (*DEFinite* or *INDefinite*). Focusing on *gettare#acqua#su_fuoco*, we find these different combinations: *gettare#DEF+acqua#su_DEF+fuoco*, *gettare#Ø+acqua#su_Ø+fuoco*, *gettare#Ø+acqua#IND+fuoco* and *gettare# Ø+acqua#su_DEF+fuoco*. Each of these articles combinations represents a specific state of the *articles entropy* for the construction at hand.

#### MODIFIERS ENTROPY

In the less syntactically frozen constructions, a slot can be modified by the presence of an adjective or a prepositional phrase attached to it. Searching for instances of *gettare#acqua#su_fuoco* in our corpus, we can find something like *gettare#molta+acqua#su_fuoco* "to minimize very much" (lit. "to throw a lot of water on the fire"), *gettare#acqua#su_fuoco+di_polemica* "to minimize the controversies" (lit. "to throw water on the fire of the controversies") or *gettare#acqua+abbondante#su_fuoco+di_insurrezione* "to abundantly minimize the uprising" (lit. "to throw abundant water on the fire of the uprising"). We use *modifiers entropy* to represent this type of variability. Noteworthily, each state in this entropy indicates the mere presence of modifying adjectives or PPs, abstracting away from the

modifier lemmas. The three types of variation for *gettare#acqua#su_fuoco* mentioned above (*gettare#molta+acqua#su_fuoco*, *gettare#acqua#su_fuoco+di_polemica* and *gettare#acqua+abbondante#su_fuoco+di_insurrezione*) are respectively represented by the states *gettare#ADJ+acqua#su_Ø+fuoco*, *gettare#Ø+acqua#su_PP+fuoco* and *gettare#ADJ+acqua#su_PP+fuoco.* Please note that we just focus on the presence of such modifying elements, their reciprocal order with the head fillers notwithstanding. As a consequence, both a preceding adjective like *molta acqua* and a following one like *acqua abbondante* are represented as *ADJ+acqua.*

**TARGET MORPHOLOGICAL ENTROPY**

This is the same as the morphological entropy above, but it's referred to the verbal target of a construction. Each verb in a SYMPAThy pattern is marked by its mood, tense, person and number. Considering all the tokens of a construction, we can see which different morphological features its verb may take and their frequencies. For *gettare#acqua#su_fuoco*, *gettare* appears as a *3rd singular indicative present* 301 times, as a *3rd singular indicative imperfect* 9 times, as a *singular masculine participle* 160 times, as an *infinitive* 359 times and so on. Each of these combinations of morphological features corresponds to a state of the *target morphological entropy* of the construction at hand.

**TARGET MODIFIERS ENTROPY**

Each verbal target can also be modified by the presence of one or more adverbs, including adverbial locutions and negations. In the case of *gettare#acqua#su_fuoco*, its verbal lemma can have, for instance, the following types of adverbial modification: *non+gettare#acqua#su_fuoco* "not to throw water on the fire", *non+gettare+mai#acqua#su_fuoco* "never to throw water on the fire", *gettare+più volte#acqua#su_fuoco* "to throw more and more water on the fire", and so on or it can have no adverbial modification. It's worth pointing out that each kind of adverb is taken into account, ranging from spatial (e.g. *qui* "here", *lì* "there", etc.) and temporal (e.g. *prima* "before", *dopo* "after", *ieri* "yesterday", etc.) ones to negations and discourse markers (e.g. *ma* "but", *però* "but", *anche* "also/too", etc.). *Target modifiers entropy* conveys this variability in the presence or absence of adverbs modifying the verbal target, but at a more abstract level than the examples just shown, which captures only the presence/absence of

modifiers and, in the second case, their number. The first and the last example above therefore count as tokens of the *gettare+MOD#acqua#su_fuoco* state, while the second one as a token of *gettare+MOD+MOD#acqua#su_fuoco*.

## 3.2. Capturing idiom semantics with distributional vectors

### *3.2.1. Distributional Semantics: Theoretical Premises*

The term *Distributional Semantics* refers to a family of approaches to semantic analysis adopted in computational linguistics and cognitive sciences that rely on the hypothesis that the degree of similarity between two linguistic expressions is a function of the similarity of the contexts in which these occur (Harris 1954; Rubenstein & Goodenough 1965; Firth 1957; Deerwester et al. 1990; Miller & Charles 1991; Lenci 2008; Sahlgren 2008; Turney & Pantel 2010). Within this perspective is therefore possible to analyze at least part of the semantic properties of linguistic expressions by studying their distributional and combinatorial properties within large corpora (Lenci 2008).

The advent of the distributional analysis is framed within the post-Bloomfieldian American structuralism and, in particular, within the studies of Harris (1951; 1954; 1970). The idea behind them is that two linguistic elements *x* and *y* that exhibit the same distribution, for example by co-occurring in the same contexts of a third element *z*, can be regarded as members of the same class (Harris 1951: 7; Sahlgren 2008: 35). According to Harris, every aspect of language can be explained via the distributional method, which thus endows the whole linguistic science with a common scientific basis (Lenci 2008). On a par with Bloomfield (1933), Harris maintains that it is impossible to analyze and fully grasp meaning in his social and extralinguistic aspects (Sahlgren 2008) and denies its role of *explanans* in linguistic research:

> As Leonard Bloomfield pointed out, if frequently happens that when we do not rest with the explanation that something is due to meaning, we discover that it has a formal regularity or 'explanation' (Harris 1970: 785).

Anyway, if this means totally excluding meaning from the horizon of linguistic research to Bloomfield (1933: 140, 145), Harris claims that meaning, in its purely linguistic aspects,

can be analyzed via the distributional methodology:

> […] if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris 1970: 786).

In this case, distributional similarity itself becomes the *explanans* for meaning. It is also interesting observing that the focus of the distributional methodology is not the *referential* meaning, but the *differential* one (Sahlgren 2006; 2008), in conformity with the Saussurean conception that "*Dans la langue, il n'y a que des differences, sans terme positif*" (Saussure 1916). In other words, the *valeur* of each sign within the *langue* cannot be defined but in opposition with the *valeur* of the other signs and the whole system is therefore based on an interaction of functional differences.

Depending on how the relationship between the distributional properties of the linguistc elements and their semantic content is conceived, a *weak distributional hypothesis* and a *strong distributional hypothesis* can be told apart (Lenci 2008).

The weak distributional hypothesis can be described as a quantitative method for semantic analysis strictly similar to the one advanced by Harris (1951). In such a perspective, the semantic properties of the linguistic expressions constrain their distribution and the relationship between the two is seen as *correlative* in nature. Thanks to this correlation, it is possible to arrive at a semantic analysis of the linguistic expressions by means of the thorough inspection of a significant number of contexts in which they occur. The classification of English verbs on the basis of their different argumental alternations proposed by Levin (1993) is compatible with the weak hypothesis, insofar as the different syntactic realizations of the argumental structures are supposed to be dictated by their semantic properties. Still within the field of lexical semantics, it has been shown that the distributional method can provide empirical confirmation to some fundamental aspects of the Generative Lexicon (Pustejovsky 1995), like coercion in English and in Italian (Pustejovsky & Jezek 2008) and verbal polysemy phenomena (Rumshisky 2008).

In his strong version, the distributional hypothesis is effectively a cognitive hypothesis. In this sense, repeated encounters with linguistic elements in different contexts have a causal role in the formation of their semantic representations and of the similarity relations between them that are stored in the mental lexicon. Such a viewpoint is firstly held in the

field of psychology by Miller and Charles (1991) under the name of *contextual hypothesis*:

> the cognitive representation of a word is some abstraction or generalization derived from the contexts that have been encountered. That is to say, a word's contextual representation is not itself a linguistic context, but is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts. The information that its contains characterizes a class of contexts (Miller & Charles 1991: 5)

This cognitive interpretation of the distributional hypothesis is, in a certain sense, a *context-based* elaboration of the Wittgensteinian motto *"The meaning of a word is its use in the language"* (Wittgenstein 1953: 20). Distributional models that rely on the strong distributional hypothesis have been employed to model a wide range of psychological phenomena, like semantic similarity judgments (Rubenstein & Goodenough 1965; Miller & Charles 1991), synonym selection (Landauer & Dumais 1997), semantic priming (Lund & Burgess 1996) and lexical acquisition (Landauer & Dumais 1997).

### *3.2.2. Vector Space Models*

The different versions of the distributional hypothesis are implemented in computational linguistics in the form of *vector space models* (*VSMs*), already used in the field of *information retrieval* (Salton et al. 1975; Sahlgren 2006; Padó & Lapata 2007; Lenci 2008; Lenci 2009; Turney & Pantel 2010). VSMs, or *Distributional Space Models* (*DSMs*), rely on a geometric metaphor of meaning, according to which words are conceived as elements in an Euclidean space and the similarity or dissimilarity between them is encoded as their proximity or distance in the space (Sahlgren 2006; Turney & Pantel 2010). The linguistic expressions under study, namely words or more complex constructions, are represented as high-dimensionality vectors. The components of such vectors capture the co-occurrence of these elements with some specific *contextual features*, which can consist of words that occur within a certain distance from the target, sentences, wider textual units and documents (Sahlgren 2006; Padó & Lapata 2007; Lenci 2008; Lenci 2009; Turney & Pantel 2010).

As Lakoff and Johnson (1980) observe, the association between *semantic similarity* and *spatial proximity* is motivated in light of the general human tendency to resort to metaphors in the conceptualization of abstract phenomena. Since our world knowledge is

first and foremost influenced by the spatio-temporal structure of our mind, the conceptual metaphors that are fundamentally used by human thought concern our physical perception of reality.

The formulation advanced by Lowe (2001) and further elaborated by Padó and Lapata (2007) represents a semantic space as a quadruple <*T*, *B*, *M*, *S*> (Lenci 2009). *T* stands for the set of target words that occur in the space; *B* is the set of elements that define the dimensions of the space and constitute the contexts employed to compare the semantic similarity between the targets; *M* is a matrix |B| x |T| that contains a vector representation of the T words, corresponding to the matrix rows, and of the B contexts in which they occur, corresponding to the matrix columns; finally, *S* stands for the similarity measure used within a given space to compute the semantic similarity between the *T* targets.

Depending on what kind of elements are selected as B and T within the quadruple, we can have *term-document*, *word-context* and *pair-pattern* matrices (Turney & Pantel 2010: 146 ff.).

Matrices of the first kind are used to compute the similarity between documents. Be **X** a term-document matrix and *m* the words in a collection of *n* documents, **X** is a matrix composed of *m* row and *n* columns. Taking $w_i$ as the *i*-th term in the vocabulary and $d_j$ as the *j*-th document in the collection, the i-th row in **X** is represented by the row vector $\mathbf{x_{i:}}$ and the j-th column is composed of the column vector $\mathbf{x_{:j}}$. Assuming that the matrix values consist of the raw occurrence frequency of the words in the documents, the $x_{ij}$ element indicates the number of times $w_j$ occurs in $d_j$. In this kind of matrices, documents are represented as unordered bags of words. Term-document matrices are firstly employed by Salton et al. (1975) in the field of information retrieval for the development of the *SMART* system. In this system, both documents and queries are rows of the same matrix. If some documents share the same column vectors with the query, they are regarded as relevant for the current search and are ranked according to the degree of similarity between their contexts and the contexts of the query vector.

The most widespread matrices in Distributional Semantics are the *word-context* ones. The context at hand typically consist of the same words used as rows (Lund & Burgess 1996; Sahlgren 2006; Bullinaria & Levy 2007; Bullinaria & Levy 2012) or in the words plus the syntactic relationship they have with the target words (Padó & Lapata 2007; Baroni & Lenci 2010). In word-word matrices, one of the most important parameters to set is the context width: a target word can be calculated as co-occuring with a given context word if, for instance, they co-occur within a five-word windows or within the same

sentence. Another parameter is the contextual direction, whereby it must be decided whether to consider as collocates of a target word just the following or preceding words or them both (Bullinaria & Levy 2007; Turney & Pantel 2010).

Considering the short example test provided by Sahlgren (2006: 69):

(72) *Die Welt ist alles was der Fall ist.*
*Was der Fall ist die Tatsache ist das Bestehen von Sachverhalten.*
*Das logische Bild der Tatsache ist der Gedanke.*

we can derive the following term-document (Table 9) and word-word matrix (Table 10):

| | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|
| Welt | 1 | 0 | 0 |
| alles | 1 | 0 | 0 |
| Fall | 1 | 1 | 0 |
| Tatsache | 0 | 1 | 1 |
| Bestehen | 0 | 1 | 0 |
| Sachverhelten | 0 | 1 | 0 |
| logische | 0 | 0 | 1 |
| Bild | 0 | 0 | 1 |
| Gedanke | 0 | 0 | 1 |

**Table 9:** term-document co-occurrence matrix (Sahlgren 2006: 69)

| | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ | $w_8$ | $w_9$ |
|---|---|---|---|---|---|---|---|---|---|
| Welt ($w_1$) | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| alles ($w_2$) | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Fall ($w_3$) | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Tatsache ($w_4$) | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| Bestehen ($w_5$) | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| Sachverhelten ($w_6$) | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| logische ($w_7$) | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| Bild ($w_8$) | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| Gedanke ($w_9$) | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

**Table 10:** word-word co-occurrence matrix (Sahlgren 2006: 70)

Lin and Pantel (2001) introduce the *extended distributional hypothesis*, which affirms that lexical patterns that co-occur with the same word pairs have the same meaning. To measure the semantic similarity between these patterns, we can create a matrix whose row vectors correspond to pairs like *mason : stone* and *carpenter:wood* and column vectors correspond to relational patterns like *"*X *cuts* Y" and "X *works with* Y". Conversely, according to the *latent relation hypothesis* (Turney 2008) it is the pairs of words co-occurring with the same patterns to exhibit the same meaning.

Importantly, the elements in a matrix can consist in mere raw co-occurrence frequency counts, Boolean values, the logarithm of the frequencies or raw frequency counts weighted by association measures, like Pointwise Mutual Information (PMI; Church & Hanks 1989) or Local Mutual Information (LMI; Evert 2008):

$$\text{(R) } PMI = log \frac{p(x,y)}{p(x)p(y)}$$

$$\text{(S) } LMI = f(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

The former association measure is more biased towards low-frequency co-occurrences, while the latter favors high frequency ones. In structured DSMs (Padó & Lapata 2007; Baroni & Lenci 2010), the contextual features vectors are composed of the two variables *r* and *w'*, the former being the syntactic relation that connects the target word *w* and the contextual word *w'* and the latter being the contextual word. Lin (1998) proposes an association measure that is specific for these cases:

$$\text{(T) } Lin(w,f) = log \frac{P(w,f)}{P(w)P(r|w)P(w'|w)}$$

Curran (2003 : 82-83) proposes to exploit a variation of one-sample t-test to calculate the associational strength between two words. One-sample t-test measures the difference between the observed mean ($\bar{x}$) and the expected mean (□) of a sample, normalized by the ratio between the variance ($s^2$) and the sample size (*N*). The null hypothesis is that the observed and the expected mean are equivalent:

$$\text{(U) } t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

When this formula is used to assess the degree of statistical association between two words, the null hypothesis is that the two words are independent. Variance can, in this case, be approximated to the expected probability *P(f)P(w)* and *N* is not considered, since the size of the reference corpus is held constant (Curran 2003: 83):

$$\text{(V) } T - test(w, f) = \frac{P(w,f) - P(w)P(f)}{\sqrt{P(f)P(w)}}$$

### *3.2.3. On semantic similarity*

According to the matrix type we choose, it is possible to capture different kinds of similarity between the target elements, whether they share the same *attributes* or the same *relations* (Medin et al. 1990; Turney 2006; Turney & Pantel 2010). Attributes and relations are distinguished in that:

> Attributes are predicates taking one argument (e.g., X is red, X is large), whereas relations are predicates taking two or more arguments (e.g., X collides with Y, X is larger than Y). Attributes are used to state properties of objects; relations express relations between objects or propositions. (Medin et al. 1990)

The *attributional similarity* between two words *a* and *b* refers to the degree of overlap between the attributes of *a* and *b*. *Dog* and *wolf* are an instance of two attributionally similar words. When two words exhibit a considerable degree of attributional similarity, they are defined *synonyms*. This kind of similarity is captured by the word-context matrices. Pair-pattern matrices, on the other hand, permit to measure the degree of *relational similarity* between two word pairs *a:b* and *c:d,* according to the degree of similarity between their relations. Pairs with a considerable degree of relational similarity are defined as *analogues.* Finally, two words are defined as *semantically associated* if they tend to co-occur frequently, like *honey* and *bee* (Chiarello et al. 1990).

A predictable criticism to the notion of semantic similarity on which DSMs rely resides in its excessive vagueness, since it comprises different kinds of semantic relationships, like synonymy, antonymy, hyponymy and hyperonymy (Sahlgren 2006; 2008). Nevertheless, a

great deal of studies have confirmed the psychological plausibility of the notion of semantic similarity. Participants to psycholinguistic experiments have been shown to spontaneously produce semantic similarity judgments between pairs of stimuli, without further specifications being provided, and showing a certain degree of inter-subject agreement (Rubenstein & Goodenough 1965; Miller & Charles 1991). Hodgson (1991) observes what kinds of semantic relationships produce lexical priming effects. Such an effect is met in all the cases in which the prime-target pair is bound by a synonymic, hyperonymic, hyponymic and antonymic relation and can therefore be associated with the general notion of semantic similarity.

Semantic similarity in a DSM is captured via a series of vector distance measures (Manning & Schütze 1999: 298 ff.; Curran 2003: 72 ff.; Sahlgren 2006: 34 ff.; Bullinaria & Levy 2007: 8 ff.; Jurafsky & Martin 2009: 697 ff.; Turney & Pantel 2010: 160 ff.).

The most common measures of distance between vectors in an Euclidean space are *Euclidean distance* (or *L2 Norm*) and *Manhattan distance* (or *Levenstein distance*, *L1 Norm*, or *City Block Distance*):

$$\text{(W) } dist_{Euclidean}(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

$$\text{(X)} dist_{Manhattan}(\vec{x},\vec{y}) = \sum_{i=1}^{N}|x_i - y_i|$$

Both measures are particular instances of the more general *Minkowski* measure:

$$\text{(Y) } dist_{Minkowski}(\vec{x},\vec{y}) = (\sum_{i=1}^{n}|x_i - y_i|^N)^{\frac{1}{N}}$$

wherein N = 2 for Euclidean distance and N = 1 for Manhattan distance.

These geometrical measures are nonetheless rarely used to calculate word similarity, for which measures coming from information retrieval are more commonly exploited. The most simple measure of vector similarity is their *scalar product*:

$$\text{(Z) } sim_{scalar\ product}(\vec{x},\vec{y}) = \vec{x}\cdot\vec{y} = \sum_{i=1}^{N} x_i y_i$$

Widdows (2004) underlines that the aforementioned measures are not optimal for the distributional models, since they show a bias towards long vectors. We know that vector

length is defined as follows:

$$\text{(AA) } |\vec{x}| = \sqrt{\sum_{i=1}^{N} x_i^2}$$

so that the bigger the component values, the longer a vector. As a consequence, frequent words will appear more distant from the other ones according to the Minkowski measures and nearer to the other ones according to the scalar product than less frequent words. The solution that is most frequently used consists in a *normalized* scalar product, that is a scalar product divided by the length of both vectors. Such a calculation corresponds to the *cosine* of the angle comprised between the two vectors and represents the most frequently used measure in the word space models:

$$\text{(AB) } sim_{cos}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2} \sqrt{\sum_{i=1}^{N} y_i^2}}$$

Cosine similarity is therefore not influenced by the vector length and returns a fixed similarity index spanning from 0, for orthogonal vectors, to 1, for identical vectors.

### *3.2.4. The problem of dimensionality reduction*

A fundamental parameter to be set in the realization of a DSM is *dimensionality*. If, on the one hand, an insufficient amount of data deprives the model of a reliable statistic basis for building semantic representations, on the other hand, the larger the amount of data, the larger the co-occurrence matrix, with a consequent reduction of the efficiency and the scalability of the algorithm (Sahlgren 2006). Another problem in DSMs is *data sparseness*, whereby the majority of the cells in a matrix are null value. This issue, which is well known within the field of computational linguistics, represents a particular case of *Zipf's law* (Zipf 1949; Manning & Schütze 1999: 23-24): just a restricted subset of all the types in a language tend to occur frequently, while the majority of them occur in a limited number of contexts. A solution to the both the problem of high dimensionality and data sparseness consists in representing the data of the high-dimensionality space in a lower-dimensionality space, while preserving, at the same time, the greatest possible amount of initial data. Starting from the 90s, an extension of the vector space models based on the

employment of *truncated Singular Value Decomposition* (*SVD*) is introduced (Deerwester et al. 1990; Landauer & Dumais 1997; Manning & Schütze 1999: 554 ff.; Turney & Pantel 2010: 159). The adoption of this model, called *Latent Semantic Indexing* (*LSI*) when applied to the measurement of the similarity between documents in the field of information retrieval (Turney & Pantel 2010), is motivated by the inability of the previous VSMs to exploit synonymy to enlarge the results of a query. Let's observe the example proposed by Manning and Schütze (1999: 554):

| | Term 1 | Term 2 | Term 3 | Term 4 |
|---|---|---|---|---|
| Query | user | interface | | |
| Document 1 | user | interface | HCI | interaction |
| Document 2 | | | HCI | interaction |

**Table 11:** an example of query (Manning & Schütze 1999: 554)

Document 1 is certainly relevant for the query at hand, since it contains its exact terms. On the other hand, also Document 2 could be relevant for it, since its terms *HCI* and *interaction* co-occur in Document 1 with the query terms. In LSI, documents and queries are projected from an initial high-dimensionality space to a lower-dimensionality one with latent semantic dimensions, wherein a query and a document can display a high cosine similarity even though they don't share any term, by both contain words that are assigned to the same latent semantic dimension. Fundamentally, the dimensions of this reduced space coincide with the axes of greatest variation in the initial space, so that the greatest amount of information is preserved. Truncated SVD applied to word similarity is called *Latent Semantic Analysis* (Landauer & Dumais 1997).

To get briefly into the mathematical details, SVD decomposes a matrix **X** into the product of three matrices $U\Sigma V^T$. U and V are in column orthonormal form, which means that the columns of U and V are orthogonal and have unit length ($U^TU = V^TV = I$), while Σ is a diagonal matrix composed of singular values. If X has rank *r*, Σ has the same rank. $\Sigma_k$, where $k < r$, will be the diagonal matrix formed from the top *k* singular values, while $U_k$ and $V_k$ will be the matrices obtained by picking the corresponding columns from U and V. $U_k\Sigma_kV^T_k$ is the *k*-rank matrix that best approximates X, by minimizing the approximation errors. More formally, $\hat{X} = U_k\Sigma_kV^T_k$ minimizes $\|\hat{X}-X\|_F$ over all matrices $\hat{X}$ of rank *k*. $\|\ldots\|_F$

stands for the Frobenius norm.

### *3.2.5. Compositionality in Distributional Semantics*

Since compositionality constitutes a fundamental component of natural language semantics, DSMs must also be able to create vector representations for complex expressions in addition to single words (Erk & Padó 2008; Mitchell & Lapata 2008; 2010; Clark 2012).

According to Mitchell and Lapata's (2010) formulation, the composition of two words *u* and *v* is a function of the two words, the syntactic relation *R* between them and additional world knowledge *K*:

$$\text{(AC) } \boldsymbol{p} = f(\boldsymbol{u}, \boldsymbol{v}, R, K)$$

The issue of compositionality has been extensively tackled within the field of Distributional Semantics, with scholars trying to find out how to represent the vector of a complex expression in the best way.

First of all, a complex expression, a phrase or a sentence could be represented as a unique vector just like they were a single word. An evident shortcoming of this method is the extreme sparseness of expressions that are longer than two words, which would result in the absence of statistically significant data. Mitchell and Lapata (2008; 2010) present an exhaustive overview of the main existing operations for vector composition in Distributional Semantics. In their exposition, they first of all set aside the *K* element, their only purpose being to study how word vectors can combine between them, without comparing the composed expressions with already existing representations. Another constraint posed by the authors is the fact that *p*, *u* and *v* must show the same dimensionality. This criterion stand at odds with other approaches that give every part of speech a different structure according to their function. Nouns are thus represented as vectors, adjectives as matrices, insofar as they behave as noun modifiers, and verbs as third-order tensors (Baroni & Zamparelli 2010; Coecke et al. 2010). Although it might sound more reasonable to assign a different dimensionality to every syntactic type, Mitchell and Lapata (2008; 2010) claim that a fixed dimensionality would make all the operations computationally less expensive and allow the comparison between different

syntactic categories within the same space.

Within the field of linear functions, we first detect an *additive* class of compositional functions, assuming that *p* is a linear function of the Cartesian product between *u* and *v*:

$$\text{(AD) } \boldsymbol{p = Au + Bv}$$

*A* and *B* are the matrices determining the contribution of *u* and *v* to *p*. Moving to the class of *multiplicative* compositional functions, we could also take *p* as a linear function of the tensor product between *u* and *v*:

$$\text{(AE) } \boldsymbol{p = Cuv}$$

*C* in (AE) is a third-rank tensor, which project the tensor product of *u* and *v* to the same space of ~ *p*. In any case, the obtained result is an approximation of a non-linear structure.

Exiting from the sector of linear multiplicative functions, we then find u-squared functions:

$$\text{(AF) } \boldsymbol{p = Duuv}$$

where D is a fourth-rank tensor. Within additive models, the most simple compositional function is *vector sum*. Given the following co-occurrence matrix:

| | music | solution | economy | craft | reasonable |
|---|---|---|---|---|---|
| practical | 0 | 6 | 2 | 10 | 4 |
| difficulty | 1 | 8 | 4 | 4 | 0 |

**Table 12:** hypothetical semantic space for *practical* and *difficulty* (Mitchell and Lapata 2010: 1401)

we would get *p* = [1 14 6 14 4]. Although vector sum is widely spread in information retrieval for the representation of wider units like sentences or documents (Landauer & Dumais 1997), it is not able to encode word order and would therefore assign the same representation to sentences containing the same words but in a different order. Vector addition, moreover, simply unites the meaning of both words, generating an intermediate element: $\overrightarrow{practical\ difficulty}$ is represented as an intermediate vector between

$\overrightarrow{practical}$ and $\overrightarrow{difficulty}$, but the actual meaning of the expression is not an intermediate one between the two components. Kintsch (2001) proposes a solution by aiming at verifying how a predicate is modified depending on the argument on which it operates (e.g. cf. *run* in *the horse ran* and *the color ran*). To vector sum, Kintsch (2001), adds also the vectors of words that are semantically close to the predicate and to the argument:

$$\text{(AG) } \boldsymbol{p} = \boldsymbol{u} + \boldsymbol{v} + \sum_i \boldsymbol{n_i}$$

Within the sum operation the *m* most similar elements to the predicate are considered, from which the *k* most similar elements to its argument are selected. As a consequence, if in the composition of *practical* and *difficulty* the similar word is *problem* (with *problem* = [2 15 7 9 1]), the final representation is composed of *practical* + *difficulty* + *problem* = [3 29 13 23 5].

An alternative is the introduction of coefficients that weight the sum components differently and make the function asymmetrical:

$$\text{(AH) } \boldsymbol{p} = \boldsymbol{\alpha v} + \boldsymbol{\beta u}$$

Going further with the same reasoning, we could even delete the contribution of one of the two elements and consider just the other one (the phrasal head, for instance):

$$\text{(AI) } \boldsymbol{p} = \boldsymbol{v}$$

In the last case, *p* would just correspond to the vector for *difficulty*. This impoverished model can act as a baseline for more complex models.

While with a sum all the vector components are added up, a multiplicative function like:

$$\text{(AL) } \boldsymbol{p} = \boldsymbol{u} \odot \boldsymbol{v}$$

takes into account just the component of *u* that are relevant for *v*. This happens because if a given component has a null value in one of the two vectors, the corresponding component in the combined vector will necessarily have a null value, so that *practical* $\odot$ *difficulty* = [0 48 8 40 0]. Although this kind of multiplication is asymmetrical, it

represents an example of the wider class of multiplicative functions, from which we can derive syntax-sensitive asymmetrical functions. Taking the function $\boldsymbol{p = Cuv}$ and considering *C* an identity matrix, we obtain the tensor product of *u* and *v*:

$$\text{(AM) } \boldsymbol{p = u \otimes v}$$

whose result is a matrix with entries $p_{ij} = u_j \cdot v_j$:

$$\text{(AN) } \overrightarrow{practical} \otimes \overrightarrow{difficulty} = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 6 & 48 & 24 & 24 & 0 \\ 2 & 16 & 8 & 8 & 0 \\ 10 & 80 & 40 & 40 & 0 \\ 4 & 32 & 16 & 16 & 0 \end{matrix}$$

Tensor product generates an array with a higher dimensionality than the constituent vectors. *Linear convolution* permits to compress the tensor product of the two vectors of *n* dimensions into a vector of *2n – 1* dimensions via the sum of the diagonal values of the tensor product matrix. In the following example by Jones and Mewhort (2007), linear convolution of two 3-dimensional vectors is performed, resulting in a final 5-dimensions *z* vector:

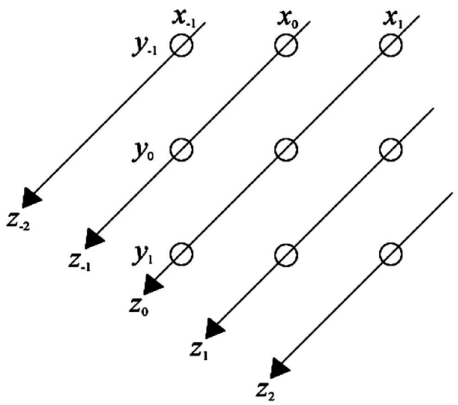



**Figure 22:** linear convolution of the two 3-dimensional vectors **x** and **y**.
The resulting vector **z** has 2·3-1=5 dimensions (Jones & Mewhort 2007: 4)

Although the final vector undergoes dimensional reduction in this way, it nonetheless remains incomparable with the initial vectors, since it has in any case a higher dimensionality. The problem can be solved via *circular convolution*, which, starting from

two *n*-dimensional vectors, generates a vector with an equal number of dimensions:

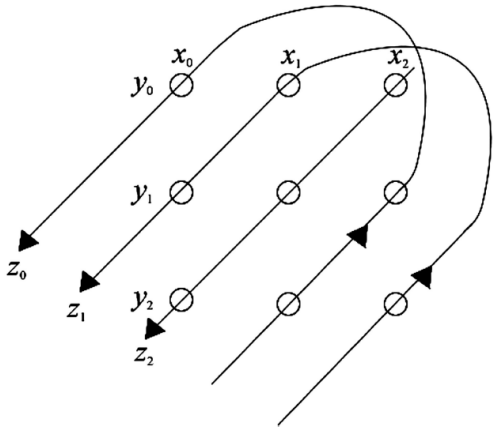



**Figure 23:** circular convolution of the two 3-dimensional vectors **x** and **y** (Jones & Mewhort 2007: 5)

In those frameworks that are grounded in formal semantics, like Montague Grammar (1974), each syntactic structure is assigned a different kind of function. In the compositional process, one of the two components acts on the other one by modifying it. These idea can be represented by resorting to the multiplicative class of functions. In the case

$$\text{(AN) } \boldsymbol{p = Cuv = Uv}$$

the product between *C* and *u* creates the matrix U, which represents a constituent that acts on the vector *v*, representing the other constituent. We previously mentioned the approach of Baroni and Zamparelli (2010), which treat adjective-noun composition representing the noun as a vector and the adjective as a matrix that maps the noun on the modified representation. Although Mitchell and Lapata (2010) represent all the syntactic categories as vectors, they employ the *Uv* product to elaborate a new compositional function based on the concept of *dilation*. Taking the simple multiplicative model $p = u \odot v$, the first constituent is multiplied by the *C* tensor generating a diagonal matrix *U* whose elements that are different from zero correspond to the components of *u*. The matrix is then multiplied by *v*, extending and shortening it in different directions. The vector is scaled by a factor $u_i$ on the i-th base. The result of this process are nonetheless dependent on the basis that is employed. To make the whole process independent from the selected base, *v* is dilated along the direction of *u* rather than the direction of the base and is decomposed in a

*x* component, parallel to *u*, and an orthogonal *y* component:

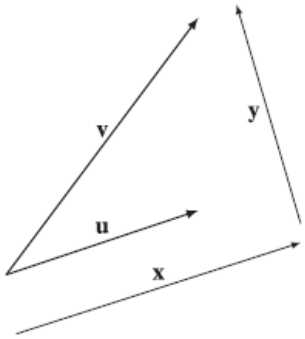



**Figure 24:** decomposition of **v** into the two components **x∥u** e **y⊥u** (Mitchell & Lapata 2010: 1405)

Since $\boldsymbol{v} = \boldsymbol{x} + \boldsymbol{y}$, inserting the scalar product on both sides of the equation, we obtain:

$$\text{(AO) } \boldsymbol{v} \cdot \boldsymbol{u} = |\boldsymbol{x}||\boldsymbol{u}| \cos(0°) + |\boldsymbol{y}||\boldsymbol{u}| \cos(90°) = |\boldsymbol{x}||\boldsymbol{u}| \rightarrow |\boldsymbol{x}| = \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{|\boldsymbol{u}|}.$$

These two vectors can be expressed in terms of *u* and *v*:

$$\text{(AP) } \boldsymbol{x} = \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{|\boldsymbol{u}|} \frac{\boldsymbol{u}}{|\boldsymbol{u}|} = \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}} \boldsymbol{u}$$

$$\text{(AQ) } \boldsymbol{y} = \boldsymbol{v} - \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{\boldsymbol{u} \cdot \boldsymbol{u}} \boldsymbol{u}$$

Dilate *x* by a factor *k*, with *y* remaining equal, we generate the modified vector *v'*, that has been stretched to emphasize the contribution of *u*:

$$\text{(AR) } \boldsymbol{v}' = \boldsymbol{\lambda x} + \boldsymbol{y} = \boldsymbol{\lambda} \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{u} \cdot \boldsymbol{u}} \boldsymbol{u} + \boldsymbol{v} - \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{u} \cdot \boldsymbol{u}} = (\boldsymbol{\lambda} - \mathbf{1}) \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{u} \cdot \boldsymbol{u}} \boldsymbol{u} + \boldsymbol{v}.$$

The expression is simplified by *u·u* in that cosine similarity is not sensitive to the vector modules:

$$\text{(AS) } p = (u \cdot u)v + (\lambda - 1)(u \cdot v)u$$

Returning to the two example vectors *practical* and *difficulty*, given the scalar products **practical x practical** = 156 and **practical x difficulty** = 96 and $\lambda$=2, we get 96 **practical** + 156 **difficulty** = [156 1824 816 1584 384].

Most importantly, Mitchell and Lapata's (2010) study confirms the cognitive

plausibility of such compositional models by performing a correlation analysis between speaker-elicited similarity judgments that are assigned to adjective-noun, noun-noun and verb-object pairs and the similarity scores calculated by the aforementioned vector combination methods. The distributional indices are calculated both from a simple semantic space and from an LDA topic model. In the case of the simple semantic space, the multiplicative scores correlate the best with the speaker judgments, followed by dilation and weighted sum. The circular convolution model, followed by the simple additive and the Kintsch model, is the one which correlates the least. In the topic model, we have the opposite situation, with additive models performing better, due to the sparseness of its representations that results in a considerable loss of information with the multiplicative functions. Still, circular convolution turns out to be the worst model.

### *3.2.6. Analyzing MWE compositionality with Distributional Semantics*

Moving from these theoretical premises, we can now ask ourselves how these vector composition measures can be profitably employed to grasp the restricted compositionality of Multiword Expressions.

Schone and Jurafsky (2001) resort to Latent Semantic Analysis to assess the non-compositionality of MWEs by measuring the cosine distance between the vector for the candidate MWE and a weighted vector sum of its constituents. They expect that the smaller the cosine, the higher the compositionality. Evaluation is carried out by comparing the extracted MWEs with those listed in dictionaries and what they find out is that these methods does not offer a significant improvement over existing methods for MWEs extraction. Anyway, as Bannard et al. (2003) note, assuming that non-compositionality should be the only requisite for the presence of a given MWE in a dictionary is not very sound from a linguistic point of view and hence it does not emerge as a reliable method for testing the MWE extracting efficacy of LSA.

Baldwin et al. (2003) too use LSA to analyze the compositionality of English noun-noun compounds and VPCs. The distributional similarity between an MWE and its head word is calculated and a correlation between similarity and compositionality is demonstrated: the higher the similarity, the higher the compositionality of the target expressions.

Venkatapathy and Joshi (2005) claim that if a multiword is highly dissimilar to its

constituent verb, the verb is not used in the multiword in its general sense. The meaning of *change* in *change hands*, for example, is quite dissimilar to its usual sense. Therefore, the more dissimilar the vector of a word combination is to the vector of its verbal component, the greater the likelihood that the expression is *de facto* a MWE. On the other hand, they remark that the verb *give* in *give a smile* has merely a function of support verb, while the meaning of the whole expression is very similar to that of the verb-form of the object, *smile.* Thus, the more similar the vector of a word combination to the vector of the verb-form of the object, the more is the likelihood that it is a MWE.

In addition to the lexical fixedness and syntactic fixedness measures also used in Fazly et al. (2009), Fazly and Stevenson (2008) make use of distributional measures to distinguish, in an unsupervised fashion, four kinds of verb-noun combinations, namely idioms, light verb constructions, abstract combinations and literals. Three distributional measures are employed: the first one, $Sim_{dist}$ *(v + n, n)*, captures the distance between a verb-noun combination and the noun constituent, the second one, $Sim_{dist}$ *(v + n, v)*, focuses on the distance between the combination and the verbal component, while the third one, $Sim_{dist}$ *(v + n, rv)*, which is especially useful for light verbs constructions, measures the cosine similarity between a word combination vector and the vector of a verb that is morphologically related to the noun component. The contextual window is composed by ± 5 words occurring from the target. In applying this third measure to the combination *make a decision*, for instance, we should calculate the similarity between the vector of *make + decision* and the vector of *decide*. What comes to the fore is that $Sim_{dist}$ *(v + n, v)* does not tell apart the different classes. This could be motivated by the fact that the employed verbs display high frequency and polysemy and thus the distributional context of such a verb may not correspond to one of its particular subsenses. $Sim_{dist}$ *(v + n, n)* separates idioms from the other three groups by assigning them the lowest degrees of compositionality. In any case, this measures does differentiate among the remaining classes. As expected, $Sim_{dist}$ *(v + n, rv)* assigns the highest scores to light verb constructions.

Křcmář et al. (2013) compare the performance in assessing the compositionality of word combinations of 5 different word space models combined with 4 different compositionality measures. The selected spaces are:

- *Vector Space Models* (*VSM*), characterized by a term-document matrix whose elements are normally weighted by the product of local and global weighting functions. Among the first group we can include no functions, logarithm + 1 and

squared root. They have the function to minimize highly occurring words in the document. Global functions may consist in no functions, inverse document frequency and entropy, with each element in the same row of the matrix being weighted by the same factor;

- *Latent Semantic Analysis* (*LSA*);
- *Hyperspace Analogue to Language* (*HAL*; Lund & Burgess 1996), which works with a word-word matrix and collects the corpus statistics by moving a double-sided window of 1 to 5 words around the target word and detecting both the following and preceding words. The result is a |V| x |2V| matrix, where |V| is the vocabulary of the corpus being analyzed, which can be reduced by keeping just the most informative columns (i.e. the column with the highest entropy $-\sum_j p_j log\, p_j$, with $p_j$denoting the probability of a word in the column *j*);
- *Correlated Occurrence Analogue to Lexical Semantics* (*COALS*) processes a corpus with a sliding window like HAL, but with a window size of 4, without distinguishing preceding and following words and discarding all the columns but the most frequent *m* representing the most common open-class words;
- *Random Indexing* (*RI*; Sahlgren 2005), which first of all assings a random vector to each word type in the corpus; each random vector is very sparse and contain several non-zero values from the {-1,1} set. Instead of accumulating the weighted co-occurrence counts of neighbouring words to the target word types, RI accumulates the random vectors of the words that co-occur with the targets.

As for the compositionality measures, we can differentiate among four types:

- *Substitutability-based measures*, which are based on the assumption that replacing a MWE component results in an anti-collocation (Pearce 2002). The compositionality of a word combination is computed as the ratio between the token frequency of the expression in a corpus and the sum of the token frequencies of its alternatives;
- *Endocentricity-based measures*, which compare the vectors of the entire expressions with those of the constituents;

- *Compositionality-based measures*, which calculate the similarity between the vectors of the expressions and the vectors resulting from the sum of product of their components;
- *Neighbors-in-common-based measure*, which considers the overlap of the most similar words to a target expression and to its constituents.

These spaces and measures are evaluated against the DISCO (Biemann & Giesbrecht 2011) and Reddy (Reddy et al. 2011) datasets, composed of adjective-noun, verb-object, subject-verb and noun-noun combinations that are manually assigned a compositionality score. All in all, LSA and COALS perform better than their basic variants (VSM and HAL), although the correlation value vary considerably according to the type of the expressions.

### *3.2.7. Our distributional measures of idiom semantics*

As we have seen, a key aspect of idioms is represented by their idiosyncratic semantics. More precisely, they differ for the identifiability of the meaning of their parts, as well as for the plausibility of a literal interpretation, in addition to the idiomatic one. While the entropic indices we described in the preceding paragraphs explore the formal behavior of an idiom, we used distributional indices to estimate how the usage of an idiom in context diverges from the prototypical usage of its components. We represented idioms and their components with vectors whose context window is constituted by the other content words appearing within the same sentence. Co-occurrences were also extracted from "la Repubblica" via SYMPAThy. We then measured the *average cosine similarity* between an idiom vector and each of its constituent word vectors. We trained two DSMs, PPMI and PLMI, by weighting co-occurrences respectively with Positive Pointwise Mutual Information and with Positive Local Mutual Information, resulting in the two distributional measures **PPMI SIMILARITY** and **PLMI SIMILARITY**. For each space we used 10000 dimensions, comprising nouns, adjectives, adverbs and verbs and excluding auxiliaries, possessives, proper nouns, negations and modals.

Noteworthily, we are not claiming that our distributional measures effectively grasp idiom compositionality, but we choose a more cautious formulation, by saying that we are comparing how different is the usage of an idiom from the prototypical usage of its

constituent words. A simple example will clarify such an observation. If a greater average similarity between the vector for *pop the question* and its constituent vectors with respect to the similarity between the vector for *kick the bucket* and its constituent would show that the former is more compositional than the latter, we must not forget that some idioms indeed contain words that never occur in isolation and whose contexts will therefore strongly resemble those of the entire expression. In this sense, an elevated average similarity between the vector for the Italian *tirare le cuoia* and its constituent vectors will not indicate compositionality, but just the fact that *cuoia* occurs just within this expression.

## 3.3. Other basic idiom statistics

Besides the measures above, we also used the following basic statistics of the idioms:

- **LOG FREQUENCY**. The logarithm of the raw frequency of an idiom;
- **LOG RELATIVE FREQUENCY**. The logarithm of the ratio between the raw frequency of an idiom and the raw frequency of its verb head. It basically corresponds to the probability to encounter an idiom given a certain verbal lexeme;
- **FIXED ARGUMENTS NUMBER**. The number of fixed (i.e. fully lexicalized) arguments of an idiom.

# CHAPTER 4
# EXPERIMENTS, RESULTS AND DISCUSSION

In this fourth chapter we describe the experiments we conducted to test the cognitive plausibility of our corpus-based measures of idiom morphosyntactic flexibility and semantic idiosyncrasy. First of all, we describe the Tabossi et al. (2011) normative data, from which we extracted the target idioms used in our study. Their work collects human-elicited ratings on a series of psycholinguistically relevant variables, such as idiom predictability, literality and syntactic flexibility. We describe the operations we conducted to extract these idioms from our corpus and to calculate the computational indices we described in the preceding chapter. After that, we present the results of the stepwise multiple regression analyses we ran by using our indices as predictors and Tabossi et al. (2011) judgments as dependent variables. This first part of the study led us to test further syntactic modifications and to collect a different kind of judgments: while Tabossi and colleagues (2011) ask subjects how the meaning of the syntactically modified version of an idiom is similar to its unmarked meaning, we just wanted to collect acceptability judgments on a 7-point scale. Thus we resorted to *CrowdFlower* (http://www.crowdflower.com/) to build a new questionnaire in which we presented the subjects with each of the 87 idioms of our study in 8 different syntactically modified versions and asked them how acceptable each sentence was from 1 to 7. Literal expressions were also included as controls in our study. We conclude by presenting and discussing the results of the t-test between the crowdsourced judgments for the literal expressions and the crowdsourced judgments for the idiomatic expressions, the results of the correlations between our acceptability judgments and the ratings by Tabossi et al. (2011) and of the stepwise multiple regression between our corpus data and our crowdsourced acceptability judgments.

## 4.1. The normative data by Tabossi and colleagues (2011)

First of all, the idiomatic expressions used in the present work were taken from a study by Tabossi and colleagues (2011), who elicited normative judgments for 245 Italian verbal idioms from a group of 740 native subjects. For each idiomatic expression, they collected

at least 40 judgments on a series of psycholinguistically relevant variables. The variables they chose to examine are the following:

- **KNOWLEDGE** – The proportion of correct meaning definitions given for each idiom. 88 participants took part in these ratings;
- **FAMILIARITY** – This judgments indicates how well the speakers thought each idiom was known by common people on a 7-point scale. 42 participants took part in these ratings;
- **AGE OF ACQUISITION** – This index indicates at what age the subjects thought they had learnt the expressions that were presented to them; this judgments were collected on a 7-point scale, structured as follows: 0–4 years (1), 5–6 years (2), 7–8 years (3), 9–10 years (4), 11–12 years (5), 13–14 years (6), 15 years or later (7); 40 participants took part in these ratings;
- **PREDICTABILITY** – The proportion of idiomatic completions given for a certain idiom, which was presented to the subjects in an example sentence and with the final word missing; 210 subjects participated to these judgments;
- **SYNTACTIC FLEXIBILITY** – Each idiom was inserted in a sentence containing one of the following five syntactic modifications: *adverb insertion*, *adjective insertion*, *left dislocation*, *passivization* and *wh-movement*. Participants evaluated how much the meaning of the idiom in the syntactically modified version was similar from 1 to 7 to its unmarked meaning, expressed in the form of a paraphrase prepared by the authors; these ratings were collected from a total of 200 speakers; importantly, syntactic flexibility judgments were first reported separately for each type of syntactic transformation (and so we have judgments regarding ADVERB INSERTION ACCEPTABILITY, PASSIVIZATION ACCEPTABILITY, etc.) and then averaged, giving rise to an overarching SYNTACTIC FLEXIBILITY variable;
- **LITERALITY** – Literality indicates the plausibility of a literal interpretation for an idiom. For instance, *perdere il treno* "to miss the boat" (lit. "to miss the train") has also a clear literal meaning beside the figurative one, while *andare in rosso* "to go into the red" does not have a plausible literal meaning and can only be idiomatically interpreted; this kind of judgments was collected on a 7-point scale from 40 subjects;

- **COMPOSITIONALITY** – Speakers were asked to evaluate how compositional each expression was from 1 to 7, that is, how much the component words of the idioms contributed to their overall meaning; 120 subjects took part in this rating.

Each idiom was also associated with a **LENGTH** value calculated in words.

Tabossi and colleagues (2011) performed a series of correlation analysis to observe whether some variables are related. The highest correlation, that captures 35% of the variance, was found between length and predictability: the longer an idiom, the more likely it is to be completed idiomatically. This finding ties in well with the results obtained by Fanari and colleagues (2010), according to which the idiomatic reading of a long string is available at its offset, while it is not available at the offset of a short idiom. Familiarity exhibits a strong inverse correlation with age of acquisition that captures 29% of the variance: the most familiar idioms are those that are made available first in the linguistic environment and are therefore acquired first. Familiarity exhibits another strong correlation with knowledge, but this is not surprising. Finally, other correlation that capture more than 10% of the variance are the one between knowledge and compositionality and the inverse one between familiarity and literality. This means that the better a speakers knows the meaning of an idiomatic expression, the more he/she sees it as compositional. On the other hand, the inverse correlation between familiarity and literality probably means that the more one is familiar with an idiom the less likely he/she is to see it as literal. In other words, the figurative meaning for a familiar idiom would be stronger than the literal one.

## 4.2. Our dataset

In the present work, we were interested to use our computational measures to model PREDICTABILITY, LITERALITY and SYNTACTIC FLEXIBILITY. From the idioms employed by Tabossi et al. (2011) we chose a subset of 87 idioms. These 87 idioms were those that occurred at least 75 times in our corpus, in order to obtain statistically reliable data. Our target expressions were extracted from the La Repubblica corpus (Baroni et al. 2004) in the form of subcategorization frames (Korhonen 2002; Schulte im Walde 2008; Lenci et al. 2012; Lenci 2014) headed by a verbal target lemma. We report in the *Appendix* the list of the idioms we used, divided into fully lexicalized (*H_lex idioms*, see below) and partially lexicalized ones (*No-H_lex idioms*).

## 4.3. Data extraction

To extract our dataset, we proceeded this way:

1. for each verbal lemma appearing in the idiom list by Tabossi et al. (2011), we extracted its SYMPAThy patterns and subcategorization frames from "la Repubblica" corpus;
2. the frames corresponding to our target idioms were identified and selected (e.g. *gettare#obj:spugna* for *gettare la spugna* 'to throw in the towel');
3. idioms with frame frequency < 75 were discarded, thereby resulting in the final dataset of 87 verbal idioms;
4. for each idiom, we calculated the entropic scores and the distributional semantic measures and the basic statistics described in Chapter 3. As regards the DSMs, we used 10000 dimensions, comprising nouns, adjectives, adverbs and verbs and excluding auxiliaries, possessives, proper nouns, negations and modals. In our dataset, we distinguished between idioms having free slots, for which we calculated also the lexical entropy (*H_Lex idioms*), and fully lexically specified ones, for which this index was not computable (*No-H_Lex idioms*)

## 4.4. First regression analysis with Tabossi et al.'s (2011) ratings

### *4.4.1. Correlational structure of our predictors*

Using our entropic, cosine and basic statistics as predictors, we ran a different stepwise multiple regression analysis for each psycholinguistic variable of interest as a dependent variable: PREDICTABILITY, LITERALITY, and SYNTACTIC FLEXIBILITY.

The dendrogram below, obtained with R (R Core Team 2015), shows the correlational structure of our predictors using Spearman's $\rho^2$ as a metric. To obtain such a dendrogram, we first extracted the correlation matrix for our predictors using Spearman's ρ. The elements of this matrix were then squared to obtain a similarity metric that was sensitive to many types of dependence, including non-monotonic relationships. Divisive hierarchical

clustering was then performed on the resulting matrix, by recursively clustering together the predictors or predictor clusters that correlated the best until we got the graph below:

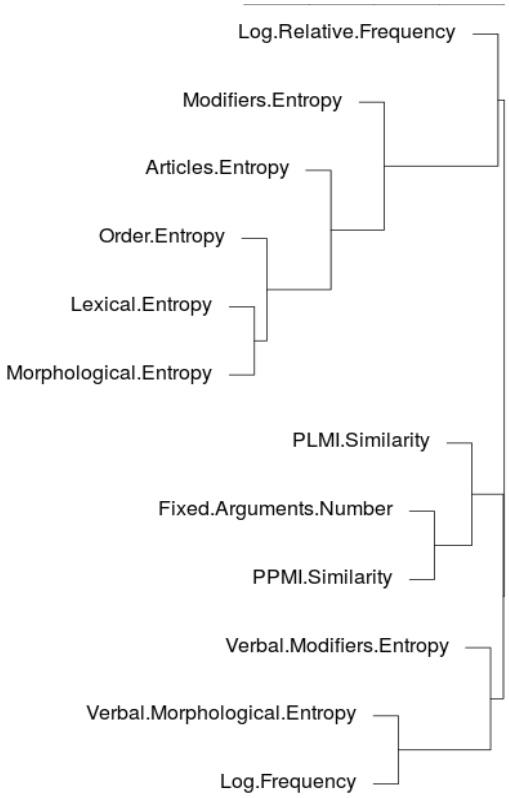



**Figure 25:** Hierarchical divisive clustering analysis for our predictors using Spearman's $\rho^2$ as a metric

The obtained clusters mirrored our theoretical assumptions. Nearly all the entropic measures clustered together, as well as the two distributional indices. The appearance of FIXED ARGUMENTS NUMBER together with the cosine measures may be due to the fact that the latter are influenced by the number of slots to be averaged. Interestingly, the verbal entropies clustered with LOG FREQUENCY. In effect, we expect that the more often an idiom occurs, the more different are the morphosyntactic contexts in which its verb appears. As we expected, LOG FREQUENCY and LOG RELATIVE FREQUENCY do not appear in the same group: while the first captures the absolute frequency of an idiom, the second one basically corresponds to the probability to encounter an idiom given a certain verbal lexeme.

### *4.4.2. Results and discussion*

As we said, using our entropic, cosine and basic statistics as predictors, we ran a

different stepwise multiple regression analysis for each psycholinguistic variable of interest as a dependent variable: PREDICTABILITY, LITERALITY, and SYNTACTIC FLEXIBILITY. All predictors were mean-centered to ensure more reliable parameter estimation, and human ratings were standardized. In our dataset, we distinguished between idioms having free slots, for which we had also calculated *lexical entropy* (*H_lex* idioms, see *Appendix*) and fully lexically specified ones, for which this index was not computable (*No H_lex* idioms). The analyses were carried out for the *H_lex idioms* and the *No H_lex idioms* separately, obtaining the six models that are described below.

The best model in each regression was chosen via the AIC criterion, which allows minor residual errors, but disadvantages the inclusion of further predictors and helps avoiding overfitting. Each final model was bootstrap-validated by 200 resampling runs.

The label assigned to each model below refers to the dataset part used (*H_lex* vs. *No H_lex* idioms) and to the modeled variable. For each model we will present a table with the *coefficient*, the *standard error* and the *t* and *p*-values for each predictor, followed by a short description of the model. The sign of the coefficient informs us about the relation between the predictor and the dependent variable: if it is positive, the higher the predictor, the higher the dependent variable; if it is negative, the higher the predictor, the lower the dependent variable. The p-value tells us whether the coefficient is significantly different from zero: a zero coefficient would mean no correlation relation between the predictor and the dependent variable. The t-value is the value of the coefficient divided by the standard error and likewise tells if a coefficient if significantly different from zero. Finally, the standard error indicates how confident we are about the estimate of the coefficient (Baayen 2008: 89-90). Predictors that were excluded or do not have significant coefficients ($p > 0.05$) in the final models will not be commented. The $R^2$ value associated with every model represents the squared correlation coefficient, *r*, and indicates, on a percentage scale or on a 0-1 range, how much of the variance was explained by the model.

**NO H_LEX PREDICTABILITY** ($F(6, 51) = 15.86$, $p = 3.52e\text{-}10$). Bootstrapped $R^2$ was 0.5792, two outliers were removed. As the coefficients show, the greater the FIXED ARGUMENTS NUMBER and the PPMI SIMILARITY between an idiom and its components, the more predictable an idiom. LOG RELATIVE FREQUENCY also received a positive coefficient

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| Fixed Arguments Number | 1.3772 | 0.2771 | 6.06 | < 0.0001 |
| PPMI Similarity | 2.7532 | 1.3188 | 2.09 | 0.0418 |
| Articles Entropy | -0.7334 | 0.4321 | -1.70 | 0.0958 |
| Modifiers Entropy | -0.6363 | 0.4754 | -1.34 | 0.1867 |
| Log Frequency | -0.1843 | 0.1140 | -1.62 | 0.1123 |
| Log Relative Frequency | 0.1635 | 0.0474 | 3.45 | 0.0011 |

**Table 13:** *No $H_{lex}$ Predictability*

**NO H_LEX LITERALITY** (F (4, 53) = 6.4, p = 0.0002806) accounted for approximately 20% of the variance of LITERALITY; two outliers were removed from the model. Both PPMI SIMILARITY and FIXED ARGUMENTS NUMBER have positive betas.

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| Fixed Arguments Number | 0.9335 | 0.2673 | 3.49 | 0.0010 |
| PPMI Similarity | 5.5378 | 1.3371 | 4.14 | 0.0001 |
| Modifiers Entropy | 0.9099 | 0.5891 | 1.54 | 0.1283 |

**Table 14:** *No $H_{lex}$ Literality*

**NO H_LEX SYNTACTIC FLEXIBILITY** (F (2, 56) = 4.296, p = 0.01838). Bias-corrected $R^2$ was 0.0713, with just one outlier excluded. The only predictor that reached significance was ARTICLES ENTROPY, which received a positive coefficient.

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| PPMI Similarity | 1.7949 | 1.2880 | 1.39 | 0.1690 |
| Articles Entropy | 1.4892 | 0.5633 | 2.64 | 0.0106 |

**Table 15:** *No $H_{lex}$ Syntactic Flexibility*

**H_LEX PREDICTABILITY** (F (6, 19) = 13.86, p = 4.802e-06) had a bootstrapped $R^2$ of 0.68, with two outliers cut out. FIXED ARGUMENTS NUMBER, PLMI SIMILARITY and ORDER

ENTROPY appeared to increase in parallel with PREDICTABILITY, while PPMI SIMILARITY, MORPHOLOGICAL ENTROPY and MODIFIERS ENTROPY showed the opposite tendency.

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| Fixed Arguments Number | 1.0871 | 0.2404 | 4.52 | 0.0002 |
| PPMI Similarity | -5.1421 | 1.2230 | -4.20 | 0.0005 |
| PLMI Similarity | 3.3898 | 1.1332 | 2.99 | 0.0075 |
| Morphological Entropy | -4.2254 | 0.8282 | -5.10 | < 0.0001 |
| Order Entropy | 6.9023 | 1.1494 | 6.01 | < 0.0001 |
| Modifiers Entropy | -10.1364 | 1.6806 | -6.03 | < 0.0001 |

**Table 16:** *$H_{lex}$ Predictability*

**H_LEX LITERALITY** (F (2, 22) = 8.965, p = 0.00142) after bootstrapping accounted for about 35% of the variance of LITERALITY. Among significant predictors, PLMI SIMILARITY obtained a positive coefficient.

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| PLMI Similarity | 8.0489 | 2.0975 | 3.84 | 0.0009 |
| Articles Entropy | -1.4271 | 0.9537 | -1.50 | 0.1488 |

**Table 17:** *$H_{lex}$ Literality*

Bootstrapped $R^2$ for **H_LEX SYNTACTIC FLEXIBILITY** (F (7, 16) = 13.02, p = 1.537e-05) was equal to 0.6875, after the removal of three outliers. Significant predictors were MORPHOLOGICAL ENTROPY, MODIFIERS ENTROPY and LOG FREQUENCY, with positive betas. Conversely, ORDER ENTROPY, ARTICLES ENTROPY, VERBAL MODIFIERS ENTROPY and LOG RELATIVE FREQUENCY had all negative coefficients.

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| Morphological Entropy | 5.7881 | 0.8633 | 6.70 | < 0.0001 |
| Order Entropy | -2.5676 | 0.9290 | -2.76 | 0.0138 |
| Articles Entropy | -3.9033 | 0.9265 | -4.21 | 0.0007 |
| Modifiers Entropy | 3.2573 | 1.4436 | 2.26 | 0.0384 |

| | | | | |
|---|---|---|---|---|
| Verbal Modifiers Entropy | -3.8651 | 0.9054 | -4.27 | 0.0006 |
| Log Frequency | 0.4351 | 0.1635 | 2.66 | 0.0171 |
| Log Relative Frequency | -0.3921 | 0.0624 | -6.28 | < 0.0001 |

**Table 18:** *$H_{lex}$ Syntactic Flexibility*

Large part of the variance in PREDICTABILITY judgments turned out to be explained by a distributional semantic representation of idioms, and by the number of their fixed arguments: the more complex an idiom and the greater the similarity between its usage and the usage of its components, the more easily subjects can predict it. Formal variability measured with entropy appeared to be relevant only for idioms with free slots, while RELATIVE FREQUENCY modeled PREDICTABILITY only for lexically specified idioms.

LITERALITY was accounted for by distributional semantic similarity indices and, for lexically rigid idioms, by the number of fixed arguments too. The portion of predicted variance (about 35% for *H_lex* and 20% for *No H_lex* idioms) was however smaller than for PREDICTABILITY. Further improvements can be expected by a better tuning of the DSMs parameters as well as by testing more advanced methods to estimate the semantic proximity between idioms and their components.

As for SYNTACTIC FLEXIBILITY, our model for lexically specified idioms explained just a restricted part of the variance (about 7%), with information on articles variability as our only significant predictor. Results were instead more promising for idioms with free slots: 68% of the variance was predicted by almost all our surface measures.

All in all, psycholinguistic ratings on idiomaticity appeared to be predictable by means of corpus-driven information that captured idiom distributional semantics, surface flexibility, frequency and the number of fully lexicalized arguments of idioms. Formal flexibility was particularly relevant for idioms with lexically underspecified slots. Interestingly, while Wulff (2009) found that parameters related to the morphological variability of the idiom verbal head had the highest weight in predicting idiomaticity judgments, VERBAL MORPHOLOGICAL ENTROPY never appeared as a significant predictor in our models. It must be noted, however, that Wulff (2009) predicted idiomaticity ratings assigned to a set of literal and figurative V-NP constructions, while the judgments we modeled only concerned idiomatic expressions. We can therefore speculate that morphological variability of the verbal head is only relevant to discriminate idiomatic vs. non-idiomatic expressions.

## 4.5. Crowdsourcing syntactic flexibility judgments

### *4.5.1. Research questions and methodological premises*

As we said above, the analysis of the data by Tabossi and colleagues (2011) and the regression analyses performed so far encouraged us to test on subjects a wider array of syntactic modifications which emerged as significant in the literature on idioms (Fraser 1970; Ernst 1981; Bianchi 1993; Nunberg et al. 1994). First of all, we wanted to distinguish internal and external modification (Ernst 1981), the former modifying just a part of an idiom (e.g. *gettare acqua sul fuoco delle polemiche*) and the latter modifying the expression as a whole (e.g. *gettare proverbiale acqua sul fuoco*, which acts as a sort of metalinguistic comment on the entire expression). Secondly, we wanted to analyze how acceptable the subjects would judge the inversion of the idiom arguments for idioms with more than one argument (e.g. *gettare acqua sul fuoco* vs. *gettare sul fuoco acqua*). Finally, we intended to insert into the test the base form of the idiom too, i.e. the idiom without any kind of syntactic modification.

From the methodological point of view, we also decided to insert literal sentences within the test, which would act as distractors. What we wanted to observe was also whether the syntactic acceptability judgments differed significantly between the idiomatic and the literal stimuli for the dimensions of syntactic variation we selected.

Finally, we said that Tabossi and colleagues (2011) asked subjects to indicate on a 7-point scale how similar the meaning of the syntactically modified version of an idiom was to its unmarked meaning, expressed in the form of a paraphrase. For example, subjects presented with the sentence *La corda venne tagliata da Gianni* "The rope was cut by Gianni" were asked how similar it was to the meaning "to flee". We thought that it would have been interesting to observe just how acceptable the subjects would find a sentence like *La corda venne tagliata da Gianni* on a 7-point scale, without any comparison to the unmarked meaning of the idiom.

Starting from these research questions, we prepared a new questionnaire and submitted it via the crowdsourcing platform *CrowdFlower* (http://www.crowdflower.com). We felt confident in resorting to crowdsourcing for collecting acceptability judgments in the light of what scholars like Sprouse (2011) affirm about the reliability of such methods. During

the last 50 years, the predominant method for studying the properties of some syntactic representation was to collect acceptability judgments as proxies for grammatical wellformedness judgments via informal experiments. In these experiments, a handful of sentences were presented to a few contributors (usually the author's colleagues) and the task took only a few minutes to be completed. In the past 15 years, some scholars have begun to exploit more formal experimental methods, like large-scale surveys, that emerge as more statistically reliable (Bard et al. 1996; Cowart 1997; Featherston 2005), but take considerably more time for their creation, for the recruitment of a sufficiently large number of subjects and for their completion. Crowdsourcing platforms such *CrowdFlower* or *Amazon Mechanical Turk* (http://www.mturk.com/) appeared as a valid alternative, since they provide rapid access to thousands of potential contributors and give the researchers all the tools necessary to submit questionnaires, collect responses and give compensations. An online interface is used to post small *Human Intelligence Tasks*, that are normally very small and simple in nature (e.g. acceptability judgments, identifying the content of images, etc.), but very numerous. Subjects are paid a small amount of money (e.g. $ 0.02 U.S.) per task and, at the end of the whole experiment, the researchers can immediately download the results in CSV format. While crowdsourcing methods have been extensively used for corpora annotation and evaluation (cfr. the proceedings of NAACL HLT 2010), their validity for syntactic acceptability tasks is still brought into question by many scholars. Differently from laboratory-based experiments, the experimenters cannot make sure that all participants belong to the required population (e.g. that they effectively are native speakers of the requested language), are not distracted, perform the task in a proper way and fully understand it. To shed some light on the issue, Sprouse (2011) compares the data of a large-scale laboratory-based syntactic acceptability experiment (176 participants) and an identical crowdsourced experiment (176 participants) as regards time, cost, participant rejection rate and differences in the shapes of the distributions of ratings for each condition. What results is that crowsourcing is a trustworthy alternative to laboratory experiments, in that it provides considerable time savings, with only a reduced participant rejection rate (i.e. less than 15%, which is within the normal boundaries for behavioral experiments). Furthermore, there is no evidence of a meaningful power loss for syntactic phenomena, nor of considerable divergence in the shapes and locations of the judgment distributions. On the other hand, crowdsourcing necessarily requires a more or less considerable amount of money, depending on the complexity of the task and the number of subjects that are required. Moreover, there is no way to control whether the participants

actually understand the task or to debrief them afterwards about potential misunderstandings with the design, the instructions and the like. Another shortcoming is that reaction times cannot be measured and that there is no way, on Amazon Mechanical Turk specifically, to automatically randomize the order of the sentences that will be presented to the subjects, in order to avoid a potential order effect in the results.

In preparing our survey, we carefully considered the recommendations and the shortcomings highlighted by Sprouse (2011) by inserting a number of test questions that roughly corresponded to 10% of our entire dataset, in order to control whether the subjects effectively understood the task and to possibly discard untrusted judgments. As we are about to explain in the following paragraph, each subject was firstly presented with a series of gold questions before moving to the effective sentences. These golden questions consisted in acceptability judgments of the same kind as those used in the rest of the study, but for which we had beforehand settled which answer was the right one. For example, we created idiomatic sentences that did not contain any sort of syntactic modifications and that were perfectly acceptable from a formal point of view and we expected for them a judgment spanning from 5 to 7. If subjects reported less than 70% correct golden questions, they were excluded from the results. Moreover, participants had the possibility to comment our golden questions in case of wrong answer and to express whether they agreed or not with the expected answer and why. Therefore we could at least marginally control wheter subjects had effectively understood the task at hand. We also made sure we had randomized the order of presentation of the sentences before submitting the task.

### *4.5.2. Participants*

Contributors coming from Italy or Switzerland and being fluent in Italian were required. A total of 30 subjects took part in our questionnaire, each of them receiving 10¢ as a compensation for every questionnaire page completed. CrowdFlower gives also the opportunity to select contributors coming from specific performance level crowds. These levels are composed of contributors that have proven over time to be particularly trustworthy. We selected contributors coming from the “Highest quality” crowd, who are the highest performance contributors accounting for 7% of monthly judgments and keep a high degree of accuracy across a very large range of CrowdFlower jobs.

### *4.5.3. Materials*

A total of 1145 sentences were created. For each of the 87 idioms in our dataset, we created up to 8 different sentences, corresponding to the following syntactic variants:

1. base form
2. adverb insertion
3. internal modification
4. external modification
5. left dislocation
6. passivization
7. interrogative wh-movement
8. inversion of the arguments order

Each variant was of course applied only where possible: intransitive idioms were not passivized and idioms with just one argument could not undergo argument order inversion. Taking the idiom *lasciare il campo a X* "to leave the field to sb.", here we provide an example of the sentences we prepared:

1. *Quando le elezioni si furono concluse, il vecchio presidente lasciò il campo al nuovo eletto.*
   "When the elections came to an end, the former President left the field to the elected one".
2. *Quando le elezioni si furono concluse, il vecchio presidente lasciò tristemente il campo al nuovo eletto.*
   "When the elections came to an end, the former President sadly left the field to the elected one".
3. *Quando le elezioni si furono concluse, il vecchio presidente lasciò il proprio campo al nuovo eletto.*
   "When the elections came to an end, the former President left his field to the elected one".
4. *Quando le elezioni si furono concluse, il vecchio presidente lasciò il proverbiale campo al nuovo eletto.*
   "When the elections came to an end, the former President left the proverbial field to the elected one".

5. *Il campo Pietro l'ha lasciato a Fabio, quando si è accorto che la sua ragazza era innamorata di lui.*
   "The field, Pietro left it to Fabio, when he realized that his girlfriend was in love with him".
6. *Il campo venne lasciato da Pietro a Fabio, quando si accorse che la sua ragazza amava lui.*
   "The field was left by Pietro to Fabio, when he realized that his girlfriend was in love with him".
7. *Che campo ha lasciato Pietro a Fabio quando si è accorto che la sua ragazza è innamorata di lui?*
   "What field did Pietro leave to Fabio, when he realized that his girlfriend was in love with him?".
8. *Pietro ha lasciato a Fabio il campo quando si è accorto che la sua ragazza è innamorata di lui.*
   "Pietro left to Fabio the field when he realized that his girlfriend was in love with him".

Crucially, each idiom was paired with a corresponding literal control expression, which was composed of the same verb and characterized by the same construction (in this case, a transitive "V Obj" one). For *lasciare il campo*, we have, for instance, *lasciare la casa* "to leave the house":

1. *Elisa lasciò la casa a dei nuovi inquilini quando cambiò città.*
   "Elisa left the house to new tenants when she changed city."
2. *Elisa lasciò tristemente la casa a dei nuovi inquilini quando cambiò città.*
   "Elisa sadly left the house to new tenants when she changed city."
3. *Elisa lasciò la sua casa a dei nuovi inquilini quando cambiò città.*
   "Elisa left her house to new tenants when she changed city."
4. 
5. *La casa Elisa l'ha lasciata a dei nuovi inquilini quando ha cambiato città.*
   "The house, Elisa left it to new tenants when she changed city."
6. *La casa venne lasciata da Elisa a dei nuovi inquilini quando cambiò città.*
   "The house was left by Elisa to new tenants when she changed city."
7. *Che casa ha lasciato Elisa ai nuovi inquilini quando ha cambiato città?*
   "Which house did Elisa leave to new tenants when she changed city?"
8. *Elisa lasciò a dei nuovi inquilini la casa quando cambiò città.*

"Elisa left to new tenants the house when she changed city."

Please note that, of course, external modification could not be applied to literal strings, being a kind of modification usually feasible just for idiomatic expressions.

We disposed 6 sentences per page and collected 5 judgments per sentence.

The questionnaire also included 150 test questions, which were of the same kind as the main questions, but were formed with idioms not included in our main dataset. These questions served the purpose to verify that the subjects had effectively understood the instructions and could efficiently perform the task.

### *4.5.3. Procedure*

The task consisted in reading each sentence and assigning it an acceptability score on a 7-point scale. We required every page to be completed in more than 30 seconds, to make sure that the subjects effectively performed the task in an efficient manner. Each subject firstly completed a page composed of 6 test questions. If the he/she obtain a score of at least 70% correct responses, he/she was then presented with the effective sentences. The subjects were presented with the following instructions (here we translate them from Italian):

*"For each sentence you are about to read, indicate on a 7-point scale how much acceptable and correct it sounds to you in Italian. A '1' corresponds to 'totally unacceptable and incorrect', while a '7' stands for 'fully acceptable and correct'. Please use intermediate values to express intermediate acceptability judgments. When you express your judgments, don't focus on sentence meaning, but to whether the sentence is well constructed or not. Examples:*

- *Carla aveva il latte alle ginocchia quando sentiva i discorsi di Antonio* = The sentence is correct and acceptable (score equal to ***6*** *or* ***7***).
- *La foglia è stata mangiata da Giulia quando lei ha scoperto il nostro gioco* = The sentence doesn't sound well constructed and acceptable (score ***from 1 to 3***).

*Thank you for your collaboration*!"

### *4.5.4. Results and discussion*

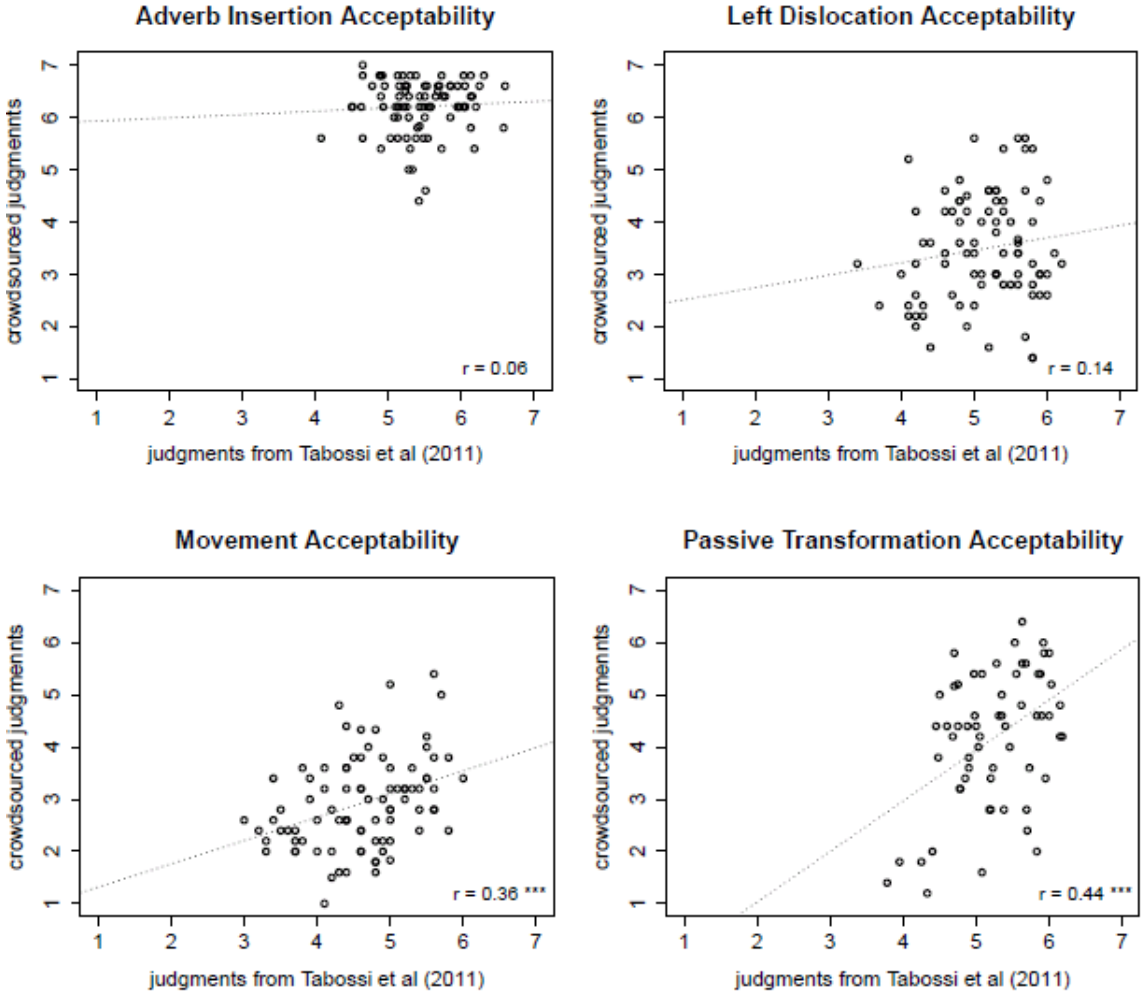
Similarly to Tabossi and colleagues (2011), we first collected judgments separately for each dimension of syntactic variation (e.g. ADVERB INSERTION ACCEPTABILITY, LEFT DISLOCATION ACCEPTABILITY, PASSIVIZATION ACCEPTABILITY) and then averaged them in order to obtain an all-embracing SYNTACTIC FLEXBILITY variable. After the questionnaire was completed, we first of all performed a Welch two sample t-test between the judgments of the idiomatic strings and the judgments given to the literal expressions for all the eights dimensions of syntactic variation tested. What came to the fore was that the acceptability judgments for each variational dimension were significantly different between the idiomatic and the literal expressions except for ADVERB INSERTION, which also according to the literature (Fraser 1970; Bianchi 1993) is the modification that idioms generally tolerate the best.

| **Compared dimension** | **t-value** | **p-value** |
|---|---|---|
| Adjective Insertion | 7.926059 | 0 |
| Adverb Insertion | 0.881097 | 0.37953 |
| Base Form | 2.259983 | 0.025113 |
| Arguments Inversion | 4.198992 | 9.1e-05 |
| Left Dislocation | 4.422076 | 1.7e-05 |
| Movement | 7.00671 | 0 |
| Passivization | 4.496181 | 1.6e-05 |
| Syntactic Flexibility | 9.6434 | 2.2e-16 |

**Table 19:** results of Welch two sample t-test between the acceptability judgments for idioms and the acceptability judgments for literal expressions that we collected via CrowdFlower, for each dimension of syntactic variation we tested.

To test whether the acceptability judgments we collected correlated or not with those elicited by Tabossi and colleagues (2011), we performed a series of correlation analyses between them, for each dimension of syntactic variation that was considered in both the studies (therefore excluding adjectival modification, since Tabossi et al. didn't distinguish between external and internal modificaition, and order inversion). The results are expressed in terms of Pearson's correlation coefficient *r*. The asterisks beside the values

indicate their level of statistical significance (*** = p < 0.001, ** = p < 0.01, * = p < 0.05). As we can observe, the correlation for the ADVERB INSERTION ACCEPTABILITY and the LEFT DISLOCATION ACCEPTABILITY judgments was not significant. Among the significant correlations, the scores on PASSIVE TRANSFORMATION ACCEPTABILITY reported the highest coefficient (r = 0.44), followed by MOVEMENT ACCEPTABILITY (r = 0.36) and SYNTACTIC FLEXIBILITY (r = 0.27). All in all, the correlational values were quite low, which confirmed that our procedure of rankings collection gave rise to different values with respect to Tabossi et al.'s (2011) work and made a second regression analysis with our new scores as dependent variables worth a try.



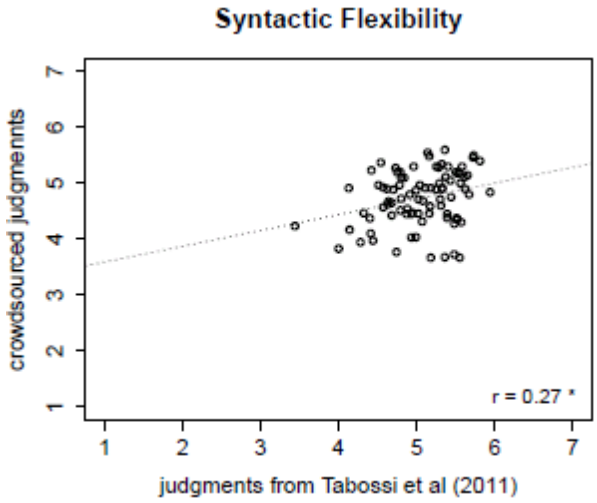



**Figure 26:** correlation graphs between the human ratings by Tabossi et al. (2011) and the ratings we collected via CrowdFlower, for each dimension of variation we tested. Pearson's *r* is used as coefficient (*** = p < 0.001, ** = p < 0.01, * = p < 0.05)

## 4.6. Second regression analysis with our crowdsourced data

Finally, we performed another series of stepwise multiple regression analyses, this time using our CrowdFlower SYNTACTIC FLEXIBILITY judgments as dependent variables. Once more, predictors were mean-centered and human ratings standardized, the best model was chosen via the AIC criterion and each final model was bootstrap-validated by 200 resampling runs. Predictors that were excluded or do not have significant coefficients ($p > 0.05$) in the final models will not be commented.

### *4.6.1. Results and discussion*

Bootstrapped $R^2$ for **NO-H_LEX SYNTACTIC FLEXIBILITY** ($F(3, 53) = 6.809$, $p = 0.0005753$) was equal to 0.2046. 3 outliers were removed. Significant predictors were ORDER ENTROPY, with a negative coefficient, and MODIFIERS ENTROPY, with a positive one.

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| PLMI Similarity | 1.7122 | 1.2432 | 1.38 | 0.1742 |
| Order Entropy | -3.5599 | 0.9473 | -3.76 | 0.0004 |
| Modifiers Entropy | 1.5373 | 0.5906 | 2.60 | 0.0120 |

**Table 20:** *No-$H_{lex}$ Syntactic Flexibility (CrowdFlower judgments)*

Bootstrapped $R^2$ for **H_LEX SYNTACTIC FLEXIBILITY** (F (7, 16) = 6.98, p = 0.0006572) resulted equal to 0.4490, with 3 outliers cut out. Significant predictors were FIXED ARGUMENTS NUMBER, PPMI SIMILARITY and MORPHOLOGICAL ENTROPY, with positive coefficients, and ORDER ENTROPY, with a negative beta.

| Predictor | Beta | S.E. | t | p-value |
|---|---|---|---|---|
| Fixed Arguments Number | 1.2717 | 0.2803 | 4.54 | 0.0003 |
| PPMI Similarity | 5.5085 | 2.1499 | 2.56 | 0.0209 |
| Morphological Entropy | 3.1908 | 1.1418 | 2.79 | 0.0130 |
| Order Entropy | -4.7486 | 1.4955 | -3.18 | 0.0059 |
| Verbal Modifiers Entropy | -1.6907 | 1.0408 | -1.62 | 0.1238 |
| Log Relative Frequency | -0.6880 | 0.3286 | -2.09 | 0.0526 |
| Verbal Morphological Entropy | 2.8713 | 2.3220 | 1.24 | 0.2341 |

**Table 21:** *$H_{lex}$ Syntactic Flexibility (CrowdFlower judgments)*

This second regression analysis showed that the indices calculated on the idioms with lexically free slots actually explained a bigger portion of the variance in the acceptability judgments with respect to the indices calculated on fully fixed idioms, in a similar fashion to what emerged from the previous regression analyses. Focusing first on fully lexically specified idioms, the more an expression can be modified by intervening adjectives and PPs, the more flexible it appeared to be perceived by subjects. Moving to *H_lex idioms*, the longer and the more morphologically variable an idiom, the more the subjects consider its various syntactic variations acceptable. Similarly, the nearer is the distributional behavior of an idiom to the prototypical behavior of its components, the more prone are the subjects to see it as flexible. Interestingly, ORDER ENTROPY exhibits an interesting behavior in both the final models: apparently, the participants tend to rate as more variable an idiom that tends to appear with its components in a fixed order.

Generally speaking, a great deal of our corpus statistics turned out to be useful in predicting human-elicited acceptability ratings on idiom flexibility. In particular, formal flexibility measures were relevant for both lexically fixed idioms and partially unspecified ones, while distributional semantic indices and the number of fixed arguments were relevant only for the second group of idioms. The polarity of some predictors (e.g. ORDER

ENTROPY) is yet to be further investigated and motivated by future contributions, but we must in any case clarify that the present analysis was just exploratory in nature and aimed at assessing the cognitive plausibility of a variety of computational techniques and measurements. This cognitive plausibility seems indeed to be confirmed by our regression study.

# CONCLUSIONS

The present thesis aimed at verifying the cognitive plausibility of computational indices capturing the formal flexibility and the semantic idiosyncrasy of a sample of Italian idiomatic expressions. The 87 idioms in our dataset were taken from the study of Tabossi and colleagues (2011), who elicited normative judgments on 245 Italian idioms from 740 native subjects on a series of psycholinguistically relevant variables, including PREDICTABILITY, LITERALITY and SYNTACTIC FLEXIBILITY.

In *Chapter 1* we observed that, notwithstanding the received equation between human language faculty and creativity and the traditional conception of single words as the fundamental units at the center of this process (Chomsky 1957; 1965; 1980; Pinker 1995), an integral part of our written and spoken production is actually composed of *Multiword Expressions* (Zgusta 1967; Erman & Warren 2000; Sag et al. 2001; Calzolari et al. 2002; Masini 2012; Siyanova-Chanturia & Martinez 2014). These can be defined as *"sequences of words acting as single units at some level of linguistic analysis"* (Calzolari et al. 2002) and encompass *collocations*, *light verb constructions*, *irreversible binomials*, *quotes* and *idioms.* The last ones are mainly characterized by non-compositionality, restricted formal flexibility, figurativity and proverbiality (Cacciari & Glucksberg 1991; Nunberg et al. 1994) and due to their challenging nature for every model of grammar (Jackendoff 1997; Goldberg 2006) have been subject to a considerable amount of psycholinguistic and neurolinguistic studies (Cacciari & Papagno 2012; Cacciari 2014). Although generative studies have for long equated idiom with non-compositionality (Katz & Postal 1963; Weinreich 1969; Fraser 1970; Chomsky 1980), Nunberg and colleagues (Nunberg 1978; Wasow et al. 1984; Nunberg et al. 1994) suggest that some idioms actually possess a certain degree of *semantic decomposability*, in that their components metaphorically map to parts of the idiomatic reference (e.g. in *spill the beans*, *spill* figuratively means "to divulge" and *beans* means "secrets") and that decomposable idioms are more morphosyntactically variable than non-analyzable ones (e.g. *the beans were spilled* vs. **the bucket was kicked*). Following contributions have nonetheless stressed that almost all kinds of idioms are formally flexible if an appropriate context is provided (Cacciari & Glucksberg 1991; Holsinger 2013; Vietri 2014). From the computational viewpoint, scholars have been mainly concerned with *type* and *token identification*, the former consisting in separating potentially idiomatic constructions (e.g. *spill the beans*) from constructions that can only have a literal meaning (e.g. *write a letter*) (Lin 1999; McCarthy

et al. 2003; Baldwin et al. 2003; Evert et al. 2004; Venkatapathy & Joshi 2005; Ritz & Heid 2006; Fazly et al. 2009) and the latter consisting in telling apart idiomatic and literal usages of a given idiomatic expression in context (e.g. *The old man kicked the bucket* vs. *Entering the junk room, I accidentally kicked a metal bucket*) (Katz & Giesbrecht 2006; Birke & Sarkar 2006; Diab & Krishna 2009; Fazly et al. 2009; Li et al. 2010; Peng et al. 2014). A reduced number of studies has addressed the question whether corpus-based idiom flexibility indices can tally with speaker-elicited idiomaticity judgments (Wulff 2008; 2009), in a similar fashion to the present study and found out that parameters related to the verbal morphology have the greatest weight in predicting the human idiomaticity ratings.

In *Chapter 2*, we began by giving a definition of *Word Combinations*, which comprise both MWEs and the preferred distributional interactions of a word, such as the argument structure constructions it typically occurs in (Lenci et al. 2014; 2015). MWEs like idioms and argument structure constructions are two examples of *constructions*, namely conventionalized pairings of form and meaning that are organized in a network called *constructicon* according to a group of approaches that go under the label of *constructionist theories* (Fillmore et al. 1988; Goldberg 1995; 2006; Croft & Cruse 2004; Hoffmann & Trousdale 2013). Constructions are organized in the constructicon according to their degree of complexity and schematicity (i.e. the inverse of lexical specification) and span from single and complex words, compounds and lexically specified idioms to partially unspecified idioms, formal idioms (i.e. *the* X-*er the* Y-*er*) and abstract syntactic constructions such as the passive or the transitive one. A considerable part of these theories adopt a *usage-based* perspective on language (Langacker 1987; Hopper 1987; Barlow & Kemmer 2000; Lieven et al. 2003; Tomasello 2003; Bybee 2006; 2010; 2013; Goldberg 2006; 2013) and claim that domain-general processes like *categorization*, *chunking*, *induction*, *cross-modal association* and *neuromotor automation* play a fundamental role in the emergence of constructions in the minds of the speakers. These Word Combinations can be computationally addressed via more constrained, POS-based methods (*P-based methods*) (Ramisch et al. 2008; 2010; Nissim & Zaninello 2013; Nissim et al. 2014; Squillante 2014) or more abstract, syntax-based ones (*S-based methods*) (Lin 1998; Blaheta & Johnson 2001; Goldman et al. 2001; Pearce 2002; Korhonen 2002; Schulte im Walde 2008; Erk et al. 2010; Seretan 2011; Lenci et al. 2012; 2014; 2015). The former are more suitable for relatively fixed, adjacent, and short combinations but miss higher level generalizations, such as the subcategorization frames most typically associated with a

given lemma, the fillers typically associated with a frame slot and the ontological classes normally occurring in such contexts. The latter, by contrast, are more prone to derive abstract generalizations like the one just described, but do not pay much attention to how words are combined at the surface level: the difference between a literal string like *non vedere l'uscita* "not to see the exit" and a syntactically identical but idiomatic one like *non vedere l'ora* "to look forward to something" can only be captured by a POS-based method, which, focusing on the surface pattern "Neg V Det N" would detect a stronger association between the words in the second string. To overcome such limitations, Lenci et al. (2014; 2015) devise *SYMPAThy* (*Syntactically Marked PATterns*), a format of data representation that unites P-based and S-based information to extract constructions from corpora.

Starting from the study by Tabossi and colleagues (2011), we extracted from the "La Repubblica" corpus (Baroni et al. 2004), a subset of 87 of the idioms they use by means of SYMPAThy and associated each of them with a *lexicosyntactic variational profile*, which comprehended:

- the variability of the fillers that instantiate the lexically free slots of the idiom;
- the morphological variability of the verb and of the idiom arguments;
- the variability of the arguments definiteness;
- the variability in the presence of adjectives and PPs modifying
- the slots, or adverbs modifying the verb;
- the variability in the linear order of the slots with respect to the verb.

As we reported in *Chapter 3*, each of these variational differences was captured via *Shannon entropy* (Shannon 1948), which measures the average uncertainty in a random variable *X*:

$$\text{(Q) } H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

and had already been exploited in acquisitional studies on syntactic productivity (Matthews & Bannard 2010) and corpus studies on idiom syntactic variability (Wulff 2008).

To computationally treat the semantic idiosyncrasy of each idiom, we resorted to Distributional Semantics (Lenci 2008; Turney & Pantel 2010), which represent the content of lexemes with vectors containing their distributional statistics with linguistic contexts. Our distributional measures calculated the average cosine similarity between the vector of

the entire idioms and the vector of each of their components, in order to get an estimate of how the usage of an idiom in context would diverge from the prototypical usage of its components.

Finally, a more basic set of statistical measures were used in our models, namely the number of fixed arguments displayed by each idiom, the logarithm of its frequency and the logarithm of its relative frequency, that is, the logarithm of the ratio between the raw frequency of an idiom and the raw frequency of its verb head. This basically corresponds to the probability to encounter an idiom given a certain verbal lexeme.

*Chapter 4* contained the results of the stepwise multilple regression analyses we ran using our computational measures as predictors and the ratings elicited by Tabossi and colleagues (2011) about idiom PREDICTABILITY, LITERALITY and SYNTACTIC FLEXIBLITY as dependent variables. According to the results, the more complex an idiom and the greater the similarity between its usage and the usage of its components, the more predictable it is. Entropic values appeared to be relevant in modeling PREDICTABILITY only for idioms with free slots, while RELATIVE FREQUENCY reached significance only for lexically specified idioms. LITERALITY was predicted to a lesser extent than PREDICTABILITY, with distributional semantic indices and the number of fixed arguments being significant. As regards SYNTACTIC FLEXIBILITY, results were better for idioms with free slots: 68% of the variance was predicted by almost all our surface indices, with respect to the 7% explained by our corpus measures for fully specified idioms. A major difference with Wulff's (2009) regression analysis caught our attention: while she finded that verbal morphology parameters had the highest weight in predicting idiomaticity judgments, VERBAL MORPHOLOGICAL ENTROPY never reached significance in our models. In any case, we must notice that Wulff (2009) predicted idiomaticity ratings assigned to a set of literal and figurative V-NP constructions, while the human ratings we used only concerned idiomatic expressions. We may therefore hypothesize that verbal morphology plays an important role only in discriminating idiomatic versus literal strings.

The first part of our analysis further encouraged us to test on speakers a wider set of syntactic modification that emerge as particularly noteworthy in the theoretical literature on idioms (Fraser 1970; Ernst 1981; Bianchi 1993; Nunberg et al. 1994). These were *internal* and *external modification*, such as in *gettare acqua sul fuoco <u>delle polemiche</u>* vs. *gettare <u>proverbiale</u> acqua sul fuoco* where the former adjective modifies just an internal part of the expression while the second one acts as a metalinguistic comment on the idiom as a whole (Ernst 1981), and *inversion of the arguments order*, like in *gettare acqua sul*

*fuoco* vs. *gettare sul fuoco acqua*. The fact that some idioms are not acceptable when its component show an inverted order, differently from literal constructions, represented an additional element of syntactic idiosyncrasy to be tested. Another interest we had was to elicit just an acceptability judgment on a 7-point scale for each syntactically modified version of our idioms, differently from Tabossi et al. (2011), who asked speakers to evaluate how similar the meaning of an idiom in a given syntactic form was to the unmarked meaning of the idiom, expressed in the form of a paraphrase. Finally, we were intentioned to pair each idiom with a corresponding literal expression, composed of the same verb and the same construction (e.g. *non vedere l'ora* was paired with *non vedere l'uscita*). Such a procedure would serve the twofold purpose of both introducing control expressions that would act as distractors for the subjects and demonstrating that the ratings were significantly different with respect to those given to the idiomatic expressions, along the syntactic variation dimensions we chose. These research questions resulted in a 1145 sentences questionnaire that was submitted via the crowdsourcing platform *CrowdFlower* (http://www.crowdflower.com). In the questionnaire, the 87 idioms of our dataset and other 87 corresponding literal expressions were inserted in up to 8 sentences, which contained them in the base form, with adverbial insertion, with external modification, with internal modification, with left dislocation, in passive form, in wh-interrogative form and with the arguments order inverted.

A t-test revealed that the scores assigned to the idioms were significantly different from those assigned to the literals with the exception of ADVERB INSERTION ACCEPTABILITY. This is not surprising, since adverbial insertion is widely recognized as the kind of syntactic transformation that idioms accept the best, similarly to literal expressions (Fraser 1970; Bianchi 1993). Interestingly, the scores elicited by Tabossi et al. (2011) turned out to be just weakly correlated to our CrowdFlower ratings, with PASSIVE TRANSFORMATION ACCEPTABILITY reporting the highest score ($r = 0.44$, $p < 0.001$) and SYNTACTIC FLEXIBILITY ($r = 0.27$, $p < 0.05$) reporting the lowest one. This led us to perform another series of stepwise multiple regression analyses with our corpus-driven measures, this time using our CrowdFlower SYNTACTIC FLEXIBILITY judgments as dependent variables. As for idioms that do not exhibit lexically free slots, the more an expression can be modified via adjective and PP insertion, the more it is perceived as flexible by the speakers. Conversely, the more variable an idiom is in the reciprocal order of its arguments, the less the subjects consider it flexible. Shifting our focus to idioms with lexically free slots, the positive relation between MORPHOLOGICAL VARIABILITY and SYNTACTIC FLEXIBILITY is confirmed,

with the addition of the NUMBER OF FIXED ARGUMENTS and the PPMI SIMILARITY between the distribution of an idiom and the distribution of its component as two relevant significant predictors of SYNTACTIC FLEXIBILITY. On the other hand, the negative relation between ORDER ENTROPY and SYNTACTIC FLEXIBILITY is confirmed.

All in all, computational indices based on Shannon entropy, Distributional Semantics, frequency and the number of arguments of each construction turned out to have psycholinguistic relevance in modeling idiomaticity. In other words, the way speakers judged the predictability, the plausibility of literal interpretation and the syntactic flexibility of a series of idiomatic expressions could be at least partly explained by our corpus-driven statistics.

The future research perspectives opened by this work are mainly twofold. In the first place, the study of the relation between computational and speaker-elicited measures of idiomaticity could be further investigated, for example by repeating our syntactic flexibility test with a greater number of judgments per sentence to observe whether consistent results with our previous analysis would be obtained or by using a wider array of distributional measures to capture idiom non-compositionality (Mitchell & Lapata 2010; Baroni 2013). Secondly, the computational techniques we applied to idioms could be used to study the lexicosyntactic productivity of a greater portion of the constructicon. Lexically fixed idiomatic expressions are located at the more flexible end of the syntactic productivity continuum, together with single words and compounds, in that they don't allow (e.g. *tagliare la corda* vs. **tagliare la fune*) or allow only restricted lexical variability (e.g. *perdere il tram/treno/autobus*). On the other hand, idioms with free slots are more located towards the middle of this continuum (e.g. *dare alla luce un progetto/un'idea/un figlio*, etc.), but there are also many other types of MWEs and Word Combinations in general that display different degrees of productivity, like light verb constructions, collocations and more abstract syntactic patterns, like the transitive or the passive constructions associated with specific lexemes. The employment of entropic, distributional and other techniques for assessing the lexicosyntactic variability of these constructions, together with the fine-tuning of efficient methods for the extraction of constructions from corpora could end up providing us, in the following years, with means to derive the entire combinatorial space of a given word starting from a corpus in a fully unsupervised way. At the same time, evidence coming from psycholinguistic and neurolinguistic studies should represent a benchmark of the cognitive validity of such computational investigations.

# APPENDIX

## A. Fully lexically specified idioms (*No-H_lex idioms*)

1. *Alzare gli occhi al cielo* “to look up to the sky”
2. *Alzare le spalle* “to shrug one’s shoulders”
3. *Andare a monte* “to come to nothing”
4. *Andare in giro* “to get about”
5. *Andare in rosso* “to go into the red”
6. *Aprire gli occhi* “to open one’s eyes”
7. *Arrivare al capolinea* “to reach the end of the line”
8. *Avere il pallino* “to have a mania for sth., to be mad for sth.”
9. *Avere voce in capitolo* “to have a voice in”
10. *Battere la fiacca* “to loaf about, to slack off”
11. *Bruciare le tappe* “to get there fast”
12. *Bussare alla porta* “to knock at the door”
13. *Cadere dal cielo* “to be heaven-sent”
14. *Cambiare le carte in tavola* “to shift one’s ground”
15. *Cantar vittoria* “to crow over one’s victory”
16. *Confondere le acque* “to muddy the waters”
17. *Essere ad un bivio* “to be at a crossroads”
18. *Essere in forma* “to bloom”
19. *Essere in gamba* “to be very capable”
20. *Fare buon viso a cattivo gioco* “to make the best of things”
21. *Fare numero* “to make up the numbers”
22. *Fare un colpo* “to carry out a raid”
23. *Farsi le ossa* “to learn the ropes, to cut one’s teeth”
24. *Gettare acqua sul fuoco* “to defuse”
25. *Gettare la maschera* “to reveal oneself”
26. *Gettare la spugna* “to throw in the towel”
27. *Giocare d’azzardo* “to gamble”
28. *Ingannare il tempo* “to while away the time”
29. *Mettere il dito sulla piaga* “to touch a sore point”

30. *Mettere la mano sul fuoco* “to stake one’s life”
31. *Mettere le carte in tavola* “to lay one’s cards on the table”
32. *Mettersi nei panni di qualcuno* “to put oneself in sb.’s shoes”
33. *Montarsi la testa* “to get a big head”
34. *Non veder l’ora* “to look forward to sth.”
35. *Pendere dalle labbra di qualcuno* “to hang off sb.’s words”
36. *Perdere il filo* “to lose the thread”
37. *Perdere il treno* “to miss the boat”
38. *Perdere la bussola* “to lose one’s head, to lose one’s bearings”
39. *Perdere la testa* “to lose one’s head”
40. *Rimboccarsi le maniche* “to roll up one’s sleeves”
41. *Rompere il ghiaccio* “to break the ice”
42. *Scagliare la prima pietra* “to cast the first stone”
43. *Scivolare su una buccia di banana* “to slip on a banana peel”
44. *Scoppiare di salute* “to be the picture of health”
45. *Scoprire l’acqua calda* “to reinvent the wheel”
46. *Soffiare sul fuoco* “to fan the flames”
47. *Tagliare la corda* “to take French leave”
48. *Tagliare la testa al toro* “to settle things once and for all”
49. *Tirare acqua al proprio mulino* “to look after number one”
50. *Tirare la corda* “to take things too far”
51. *Tirare la cinghia* “to tighten the belt”
52. *Tirare i remi in barca* “to back down”
53. *Toccare il fondo* “to reach the bottom”
54. *Togliere le castagne dal fuoco* “to pull the chestnuts out of the fire”
55. *Uscire dal tunnel* “to survive”
56. *Vedere la luce* “to see the light”
57. *Vendere cara la pelle* “to fight tooth and nail”
58. *Venire alle mani* “to come to blows”
59. *Voltare pagina* “to turn a corner”
60. *Vuotare il sacco* “to spill the beans”

## B. Idioms with lexically free slots (*H_lex idioms*)

1. *Andare a fondo di/con X* "to get to the bottom of sth."
2. *Andare a genio a X* "to sit well with sb."
3. *Appendere X al chiodo* "to hang up sth."
4. *Aprire le porte a X* "to open the floodgates to sb./sth."
5. *Avere un occhio di riguardo per X* "to have special consideration for sth./sb."
6. *Avere X sulle spalle* "to be responsible for sth./sb."
7. *Chiudere un occhio su X* "to turn a blind eye to sth."
8. *Dare carta bianca a X* "to give carte blanche to sb., to give free rein to sb."
9. *Dare del filo da torcere a X* "to give sb. a hard time"
10. *Dare X alla luce* "to give birth to sth./sb."
11. *Essere in ballo in X* "to be at stake in sth."
12. *Fare il callo a X* "to get used to sth."
13. *Lasciare il campo a X* "to leave the field to sb./sth."
14. *Leggere X tra le righe* "to read sth. between the lines"
15. *Mandare X a monte* "to foul sth. up"
16. *Mandare X al diavolo* "send sb. packing"
17. *Mettere X sul piatto della bilancia* "to weigh sth."
18. *Mostrare i denti a X* "to bare one's teeth to sb."
19. *Mozzare il fiato a X* "to take one's breath away"
20. *Passare la palla a X* "to pass the ball to sth."
21. *Passare X al setaccio* "to comb sth."
22. *Piantare X in asso* "to jilt sb., to leave sb. behind"
23. *Prendere X di petto* "to face sth. head on"
24. *Preparare il terreno a X* "to prepare the ground to sth."
25. *Sbattere la porta in faccia a X* "to slam the door in sb.'s face"
26. *Tenere X a battesimo* "to inaugurate sth."
27. *Tenere X sulla corda* "to keep on sb. on tenterhooks"

# REFERENCES


Adam, M. 1987. “A Dictionary of American Idioms”. New York: Bsl'rows Educational

Ahrens, K. 1995. “The Mental Representation of Verbs” UCSD dissertation

Aijmer, K. 1996. “Conversational Routines in English: Convention and Creativity”. London: Longman.

Altenberg, B. 1998. “On the phraseology of spoken English: The evidence of recurrent word-combinations”. na.

Altenberg, E. P. 1991. “Assessing first language vulnerability to attrition”. First language attrition, 189.

Attardi, G. and F. Dell’Orletta, 2009, “Reverse revision and linear tree combination for dependency parsing,” in Proceedings of NAACL 2009, pp. 261–264.

Baayen, H. 1992. “Quantitative aspects of morphological productivity”. In Yearbook of morphology 1991 (pp. 109-149). Springer Netherlands.

Baayen, R. H. 2008. “Analyzing linguistic data: A practical introduction to statistics using R”. Cambridge University Press.

Baldwin, T. C. Bannard, T. Tanaka, and D. Widdows, 2003, “An empirical model of multiword expression decomposability,” in Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, pp. 89–96.

Bally, C. 1951[1909], “Traité de stylistique française, Vol. I”, Ginevra, Librairie Georg and Cie S. A. and Parigi, Librairie C. Klincksieck [prima ed. Traité de stylistique française, Stoccarda, Winter, 1909].

Bannard, C. 2007. “A measure of syntactic flexibility for automatically identifying multiword expressions in corpora”. In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions (pp. 1-8). Association for Computational Linguistics.

Bannard, C. Baldwin, T. and Lascarides, A. 2003. “A statistical approach to the semantics of verb-particles”. In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18 (pp. 65-72). Association for Computational Linguistics.

Bard, E. G. Robertson, D. and Sorace, A. 1996 “Magnitude estimation of linguistic acceptability”. Language 72 (1), 32–68.

Barðdal, J. 2008. “Productivity: Evidence from case and argument structure in Icelandic”. Amsterdam and Philadelphia: John Benjamins

Bar-Hillel, Y. 1955. “Idioms. Language and information: selected essays on their theory and application”, 47-55.

Barlow, M. and Kemmer, S. (eds.). 2000. “Usage-based models of language”. Stanford:

CSLI Publications.

Baroni, M. S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, and M. Mazzoleni, 2004, “Introducing the La Repubblica corpus: A large, annotated, tei(xml)-compliant corpus of newspaper italian,” in Proceedings of LREC 2004, pp. 1771–1774.

Baroni, M. 2013. “Composition in distributional semantics”. Language and Linguistics Compass 7(10): 511-522

Baroni, M. and Lenci, A. 2010, “Distributional Memory: A general framework for corpus-based semantics”, Computational Linguistics, 36 (4), pp. 673-721

Barsalou, L. W. 1992. “Frames, concepts, and conceptual fields”

Bartsch, S. 2004. “Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence”. Gunter Narr Verlag.

Bauer, L. 1983. “English Word-formation”, Cambridge, Cambridge University Press

Becker, J. D. 1975. “The phrasal lexicon”, in Nash-Webber, B. L. and Schank, R. (eds.), “Theoretical Issues in Natural Language Processing 1”, Cambridge, Bolt Beranek and Newman Inc. 60-63.

Belletti, A. 1988. “The case of unaccusatives”. Linguistic inquiry, 1-34.

Belletti, A. 1990. “Generalised Verb Movement–Aspects of Verb Syntax”. Torino, Italy: Rosenberg and Sellier.

Berry-Rogghe, G. L. M. 1974. “Automatic identification of phrasal verbs”. In Mitchell, J. L. (ed.), “Computers in the Humanities”. Edinburgh: Edinburgh University Press, 16–26.

Bianchi, V. 1993. “An empirical contribution to the study of idiomatic expressions”. Rivista di linguistica, 5, 349-385.

Biber, D. Johansson, S. Leech, G. Conrad, S. Finegan, E. and Quirk, R. 1999. “Longman grammar of spoken and written English (Vol. 2)”. MIT Press.

Birke, J. and A. Sarkar, 2006, “A clustering approach to the nearly unsupervised recognition of nonliteral language,” in Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL06), Trento, Italy, pp. 329–336.

Blaheta, D. and Johnson, M.. 2001. “Unsupervised learning of multi-word verbs”. In 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39), pages 54–60

Bloomfield, L. 1933, “Language”. New York, Henry Holt and Company.

Bobrow, S. A. and Bell, S. M. 1973. “On catching on to idiomatic expressions”. Memory and Cognition, 1(3), 343-346.

Bolinger, D. 1976 "Meaning and memory", Forum Linguisticum I: 1-14.

Bolinger, D. L. M. 1977. “Meaning and form”. Longman Publishing Group.

Bowerman, M. 1988. “The ‘no negative evidence’ problem: How do children avoid constructing an overly general grammar?” In J. Hawkins (Ed.), Explaining language universals (pp. 73–101). Oxford: Basil Blackwell.

Bréal, M. 1904[1897], “Essai de sémantique”, Parigi, Hachette.

Bresnan, J. 1981. “An approach to Universal Grammar and the mental representation of language”. Cognition, 10(1), 39-52.

Bresnan, J. 1982. “The mental representation of grammatical relations (Vol. 1)”. The MIT Press.

Bullinaria, J.A. and Levy, J.P. 2007, “Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study”, Behavior Research Methods, 39: 510-526

Bullinaria, J.A. and Levy, J.P. 2012, “Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD”, Behavior Research Methods

Burt, J. S. 1992. “Against the lexical representation of idioms”. Canadian Journal of Psychology/Revue canadienne de psychologie, 46(4), 582.

Butt, M. 2003. “The light verb jungle”. In Harvard Working Paper in Linguistics, volume 9, pages 1–49.

Bybee, J. 1985. “Morphology: A Study of the Relation between Meaning and Form”. Typological Studies in Language 9. Amsterdam and Philadelphia: John Benjamins

Bybee, J. 1995. “Regular morphology and the lexicon. Language and cognitive processes”, 10(5), 425-455.

Bybee, J. 2002. “Sequentiality as the Basis of Constituent Structure,” in Talmy Givón and Bertram Malle (eds.), “The Evolution of Language from Pre-Language”. Amsterdam: John Benjamins, 109–32.

Bybee, J. 2003. “Mechanisms of Change in Grammaticization: The Role of Frequency,” in Brian D. Joseph and Richard D. Janda (eds.), “The Handbook of Historical Linguistics”. Oxford: Blackwell, 602–23.

Bybee, J. 2006. “From usage to grammar: The mind's response to repetition”. Language, 711-733.

Bybee, J. 2007. “Frequency of Use and the Organization of Language”. Oxford: Oxford University Press

Bybee, J. 2010. “Language, usage and cognition”. Cambridge: Cambridge University Press.

Bybee, J. 2013. “Usage-based theory and exemplar representations of constructions”. In T. A. Hoffmann and G. Trousdale (Eds.), The Oxford Handbook of Construction Grammar (pp. 49–69). Oxford: Oxford University Press.

Bybee, J. and Hopper, P. 1991. “Introduction to frequency and the emergence of linguistic structure”, Frequency and the emergence of linguistic structure, 45, 1.

Bybee, J. L. and Beckner, C. (2009). “Usage-based Theory,” in Heiko Narrog and Bernd Heine (eds.), Handbook of Linguistic Analysis. Oxford: Oxford University Press, 827–55.

Bybee, J. and Eddington, D. 2006. “A usage-based approach to Spanish verbs of becoming”. Language, 323-355.

Bybee, J. and Thompson, S. 1997. “Three frequency effects in syntax”. Berkeley Linguistics Society, 23, 65–85.

Cacciari, C. and Glucksberg, S. 1994. “Understanding figurative language”. Gernsbacher, Morton Ann (Eds.), 1994. “Handbook of psycholinguistics”, (pp. 447-477). San Diego, CA, US: Academic Press, xxii, 1174 pp.

Cacciari, C. and S. Glucksberg, 1991, “Understanding idiomatic expressions: The contribution of word meanings,” Understanding word and sentence, pp. 217–240

Cacciari, C. and Tabossi, P. 1993. “Idioms: Processing, Structure, and Interpretation”, Psychology Press

Cacciari, C. and Papagno, C. 2012. “Neuropsychological and neurophysiological correlates of idiom understanding: How many hemispheres are involved”. Neuropsychology of Language, Wiley-Blackwell, 368-385.

Cacciari, C. and Tabossi, P. 1988. “The comprehension of idioms”. Journal of memory and language, 27(6), 668-683.

Cacciari, C. 2014, “Processing multiword idiomatic strings: Many words in one?” The Mental Lexicon, vol. 9, no. 2, pp. 267–293.

Cacciari, C. Padovani, R. and Corradini, P. 2007. “Exploring the relationship between individuals’ speed of processing and their comprehension of spoken idioms”. European Journal of Cognitive Psychology, 19(3), 417-445.

Callison-Burch, C., and Dredze, M. 2010, June. “Creating speech and language data with Amazon's Mechanical Turk”. In “Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk” (pp. 1-12). Association for Computational Linguistics.

Calzolari, N. C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli, 2002, “Towards best practice for multiword expressions in computational lexicons.” in Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Canary Islands - Spain

Casadei, F. 1996. “Metafore ed espressioni idiomatiche. Uno studio semantico sull’italiano”. Roma: Bulzoni.

Chafe, W. L. 1968. “Idiomaticity as an anomaly in the Chomskyan paradigm”. Foundations of language, 109-127.

Chomsky, N. 1975. “The logical structure of linguistic theory”. Cambridge: MIT.

Chomsky, N. 1986. “Knowledge of language: Its nature, origin, and use”. Greenwood

Publishing Group.

Chomsky, N. 1991. “Some notes on economy of derivation and representation”. Principles and parameters in comparative grammar, dir. R. Freidin, 417-54

Chomsky, N. 1992. “A minimalist program for linguistic theory (= MIT Occasional Papers in Linguistics 1)”. Cambridge, Massachusetts

Chomsky, N. 2000. “New horizons in the study of language and mind”. Cambridge University Press.

Chomsky, N. 1957, “Syntactic Structures”. The Hague, Mouton.

Chomsky, N. 1959. “A review of BF Skinner's Verbal Behavior”. Language, 35(1), 26-58.

Chomsky, N. 1965, “Aspects of the Theory of Syntax”. Cambridge, MA, MIT

Chomsky, N. 1980, “Rules and Representations”. Oxford, Blackwell.

Chomsky, N. 1981, “Lectures on Government and Binding”. Dordrecht, Foris.

Chomsky, N.1966, “Cartesian Linguistics”. Harper and Row, New York.

Church, K. Hanks, P. Hindle, D. and Gale, W. 1991. “Using Statistics in Lexical Analysis”, pages 115–164. Lawrence Erlbaum.

Church, KW. Hanks, P. 1989. “Word association norms, mutual information and lexicography”. Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, University of British Columbia, Vancouver, Canada

Cinque, G. 1990. “Types of Ā-dependencies”. MIT press.

Clark, S. 2012. “Vector Space Models of Lexical Meaning”, in edited by Lappin S. & Fox, C. (eds.), Wiley-Blackwell Handbook of Contemporary Semantics — second edition.

Cover, T. M. and Thomas, J. A. 1991. “Elements of lnformation Theory”. A Wiley—Interscience Publication, New York

Cowart, W. 1997. “Experimental syntax: Applying objective methods to sentence judgments”. Thousand Oaks: Sage.

Cowie, A. P. 1988, “Stable and creative aspects of vocabulary use”, in Carter, R. and McCarthy, M. (eds.), “Vocabulary and Language Teaching”, Londra/New York, Longman, 126-139

Croft, W. and Cruse, D. A. 2004. “Cognitive linguistics”. Cambridge University Press.

Cruse, D. A. 1986. “Lexical semantics”. Cambridge University Press.

Culicover, P. W. 1976. “A constraint on coreferentiality”. Foundations of Language, 109-118.

Curran, J.R. 2003, “From Distributional to Semantic Similarity”, PhD thesis, University of Edinburgh

Cutting, J. C. and Bock, K. 1997. “That’s the way the cookie bounces: Syntactic and

semantic components of experimentally elicited idiom blends". Memory and Cognition, 25(1), 57-71.

Dell, G. S. 1986. "A spreading-activation theory of retrieval in sentence production". Psychological review, 93(3), 283.

Dell'Orletta, F. "Ensemble system for Part-of-Speech tagging," in Proceedings of EVALITA 2009.

Descartes, R. 1649/1927, "Letter (to Morus)" . In R.M. Eaton, ed. Descartes Selections

Diab, M. T. and Krishna, M. 2009. "Unsupervised classification of verb noun multi-word expression tokens". In Computational Linguistics and Intelligent Text Processing (pp. 98-110). Springer Berlin Heidelberg.

Diab, M.T. and M. Krishna, 2009, "Unsupervised classification of verb noun multi-word expression tokens," in Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '09), pp. 98–110.

Drew, P. and Holt, E. 1988. "Complainable matters: The use of idiomatic expressions in making complaints". Social Problems, 35(4), 398-417.

Dubois, J et al. 1979. "Dizionario di linguistica", I. Loi Corvetto and L. Rosiello (eds.), Bologna, Zanichelli

Elman, J. L. and Bates, E. 1997. "Response to Letters." Science 276: 1180

Erk, K. and Padó, S. 2008. "A Structured Vector Space Model for Word Meaning in Context", in Proceedings of EMNLP 2008

Erk, K. Padó, S. and Padó, U. 2010. "A flexible, corpus-driven model of regular and inverse selectional preferences". Computational Linguistics, 36(4), 723-763.

Erman, B. and Warren, B. 2000. "The idiom principle and the open choice principle". Text-Interdisciplinary Journal for the Study of Discourse, 20(1), 29-62.

Ernst, T. 1981. "Grist for the linguistic mill: Idioms and 'extra'adjectives". Journal of Linguistic Research, 1(3), 51-68.

Evert, S. 2008. "Corpora and collocations". Corpus Linguistics. An International Handbook, 2, 223-233.

Evert, S. and Krenn, B. 2001. "Methods for the qualitative evaluation of lexical association measures". In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (pp. 188-195). Association for Computational Linguistics.

Evert, S. and Krenn, B. 2005. "Using small random samples for the manual evaluation of statistical association measures". Computer Speech and Language, 19(4), 450-466.

Evert, S. 2008, "Corpora and collocations," Corpus Linguistics. An International Handbook, vol. 2, pp. 223–233.

Evert, S. U. Heid, and K. Spranger, 2004, "Identifying morphosyntactic preferences in collocations," in Proceedings of the 4th International Conference on Language Resources

and Evaluation (LREC04), Lisbon, pp. 907–910.

Fanari, R. Cacciari, C. and Tabossi, P. 2010. “The role of idiom length and context in spoken idiom comprehension”. European Journal of Cognitive Psychology, 22(3), 321-334.

Fazly, A. and S. Stevenson, 2008, “A distributional account of the semantics of multiword expressions,” Italian Journal of Linguistics, vol. 1, no. 20, pp. 157–179.

Fazly, A. Cook, P. and Stevenson, S. 2009, “Unsupervised type and token identification of idiomatic expressions,” Computational Linguistics, vol. 1, no. 35, pp. 61–103.

Featherston, S. 2005. “Magnitude estimation and what it can do for your syntax: Some wh-constraints in German”. Lingua, 115, 1525–1550.

Feldman, A. and Peng, J. 2013. “Automatic detection of idiomatic clauses”. In Computational Linguistics and Intelligent Text Processing (pp. 435-446). Springer Berlin Heidelberg.

Fellbaum, C. 1993. “The determiner in English idioms”. Idioms: Processing, structure, and interpretation, 271-296.

Fernando, C. 1996. “Idioms and Idiomaticity”. Oxford: Oxford University Press.

Fillmore, C. J. 1985. “Frames and the semantics of understanding”. Quaderni di semantica, 6(2), 222-254.

Fillmore, C. J. and Atkins, B. T. 1992. “Toward a frame-based lexicon: The semantics of RISK and its neighbors”. Frames, fields, and contrasts: New essays in semantic and lexical organization, 103.

Fillmore, C. J. Kay, P. and O'connor, M. C. 1988. “Regularity and idiomaticity in grammatical constructions: The case of let alone”. Language, 501-538.

Firth, J. R. 1957a, “A Synopsis of Linguistic Theory 1930-1955”, in AA.VV., “Studies in Linguistic Analysis”, Special volume of the Philological Society, Oxford, Blackwell, 1-32

Firth, J. R. 1957b, “Modes of meaning”, in Firth, J. R., “Papers in Linguistics 1934-1951”, Londra, Oxford University Press.

Fraser, B. 1970. “Idioms within a transformational grammar”. Foundations of language, 22-42.

Frege, G. 1892. “On concept and object.” Reprinted in B. McGuiness, ed. Collected Papers on Mathematics, Logic, and Philosophy. Oxford: Blackwell, 1984: 182–194.

Gazdar, G. 1982. “Phrase structure grammar”. In The nature of syntactic representation (pp. 131-186). Springer Netherlands.

Gibbs, R. and Nayak, N. 1989. “Psycholinguistic studies on the syntactic behavior of idioms”. Cognitive Psychology 21, 100–138

Gibbs, R. W. 1990. “Psycholinguistic studies on the conceptual basis of idiomaticity”.

Cognitive Linguistics, 1(4), 417-452.

Gibbs, R. W. and Gonzales, G. P. 1985. “Syntactic frozenness in processing and remembering idioms”. Cognition, 20(3), 243-259.

Gibbs, R. W. Nayak, N. P. and Cutting, C. 1989. “How to kick the bucket and not decompose: Analyzability and idiom processing”. Journal of memory and language, 28(5), 576-593.

Givón, T. 1979. “On Understanding Grammar”. New York and San Francisco: Academic Press.

Glucksberg, S. 1993. “Idiom meanings and allusional content”. In Cacciari, C. and Tabossi, P. “Idioms: Processing, Structure, and Interpretation”. LEA, Mahwah, NJ, pages 3–26.

Glucksberg, S. 2001. “Understanding figurative language: From metaphor to idioms”. Oxford University Press

Glucksberg, S. and Keysar, B. 1990. Understanding metaphorical comparisons: Beyond similarity. Psychological review, 97(1), 3.

Goldberg, A. E. 2006. “Constructions at work: The nature of generalization in language”. Oxford University Press.

Goldberg, A. E. 2013. “Constructionist approaches”. The Oxford handbook of construction grammar, 15-31.

Goldberg, A. E. 1995. “Constructions. A Construction Grammar Approach to Argument Structure”. Chicago: University of Chicago Press.

Goldman, J.P. Nerima, L. and Wehrli, E. 2001. Collocation extraction using a syntactic parser. In 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39), pages 61–66

Graffi, G. 1991, “La sintassi tra Ottocento e Novecento”, Bologna, Il Mulino.

Graliński, F. 2012. “Mining the Web for Idiomatic Expressions Using Metalinguistic Markers”. In Text, Speech and Dialogue (pp. 112-118). Springer Berlin Heidelberg.

Grant, L. E. 2005. “Frequency of ‘core idioms’ in the British National Corpus (BNC)”. International Journal of Corpus Linguistics, 10(4):429–451

Greenberg, J. H. 1963. “Some universals of grammar with particular reference to the order of meaningful elements”. Universals of language, 2, 73-113.

Gregoire, N. 2010. “DuELME: a Dutch electronic lexicon of multiword expressions”. Lang. Resources Eval. 44, 1–2, 23–39.

Gries, S. T. 2008. “Phraseology and linguistic theory: A brief survey. Phraseology: An interdisciplinary perspective”, 3-25.

Gropen, J. Pinker, S. Hollander, M. Goldberg, R. and Wilson, R. 1989. “The learnability and acquisition of the dative alternation in English”. Language, 65(2), 203–

257.

Hanks, P. 2013. “Lexical Analysis: Norms and Exploitations”. MIT Press, Cambridge, MA.

Hanks, P. and Pustejovsky, J. 2005. “A pattern dictionary for natural language processing”. Revue Française de linguistique appliquée, 10(2), 63-82.

Harris, Z. S. 1951. “Methods in structural linguistics”. University of Chicago Press.

Harris, Z. S. 1954. “Distributional structure”. Word, Vol. 10. 146-162

Hashimoto, C. and Kawahara, D. 2008. “Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features”. In Proceedings of the conference on empirical methods in natural language processing (pp. 992-1001). Association for Computational Linguistics.

Hashimoto, C. Sato, S. and Utsuro, T. 2006. “Japanese idiom recognition: Drawing a line between literal and idiomatic meanings”. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 353-360). Association for Computational Linguistics.

Hearst, M. A. 1992. “Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics-Volume 2” (pp. 539-545). Association for Computational Linguistics.

Higginbotham, J. 1987. “Indefiniteness and predication”. The representation of (in) definiteness, 14.

Hoffmann, T. and G. Trousdale, Eds. The Oxford Handbook of Construction Grammar. Oxford: Oxford University Press, 2013.

Hoffmann, T. and Trousdale, G. editors. 2013. “The Oxford Handbook of Construction Grammar”. Oxford University Press, Oxford.

Holsinger, E. 2013. “Representing idioms: Syntactic and contextual effects on idiom processing”. Language and Speech 56(3), 373–394.

Holsinger, E. and Kaiser, E. 2013. “Processing (non) compositional expressions: Mistakes and recovery”. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(3), 866.

Hopper, P. 1987. “Emergent grammar”. Berkeley Linguistic Society, 13, 139–157.

Hopper, P. 1988. “The APGP And The Emergence Of Grammar”. Linguistics in Context: Connecting Observation and Understanding: Lectures from the 1985 LSA/TESOL and NEH Institutes, 29, 117.

Howarth, P. 1998. “The phraseology of learners’ academic writing. Phraseology: Theory, analysis, and applications”, 161-186.

Hudson, J. A. 1990. “Constructive processing in children's event memory”. Developmental Psychology, 26(2), 180.

Hughlings Jackson, J. 1874. “On the nature of the duality of the brain”. In J.Taylor (ed.)

1958. Selected writings of John Hughlings Jackson, vol. 2. London: Staples Press, 129–145.

Humboldt, W. V. 1988[1836]. “On language”. Trans. Peter Heath. Cambridge, Eng.: Cambridge UP.

Israel, M. 1996. “The way constructions grow”. In Adele E. Goldberg (ed.), Conceptual structure, discourse and language, 217–230. Stanford, CA: CSLI Publications

Jackendoff, R. 1995. “The Boundaries of the Lexicon”, in M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, eds. “Idioms: Structural and Psychological Perspectives”, 133-165. Hillsdale, NJ: Erlbaum

Jackendoff, R. 1997, “The Architecture of the Language Faculty”. MIT Press.

Jackendoff, R. and Pinker, S. 2005. “The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky)”. Cognition, 97(2), 211-225.

Jespersen, O. 1976[1924]. “Living grammar”. In “The philosophy of grammar”. London: George Allen and Unwin, 17–29. Reprinted in D.D. Bornstein (ed.) 1976. “Readings in the theory of grammar”. Cambridge, MA:Winthrop Publishers, 82–93.

Kaschak, M. P. and Glenberg, A. M. 2000. “Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension”. Journal of memory and language, 43(3), 508-529.

Katz, G. and E. Giesbrecht, 2006, “Automatic identification of noncompositional multi-word expressions using latent semantic analysis” in Proceedings of the ACL06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, pp. 12–16.

Katz, J. 1973. “Compositionality, idiomaticity, and lexical substitution”. In: Anderson, S. Kiparsky, P. (Eds.), “A festschrift for Morris Halle”. Holt, New York, pp. 357-376.

Katz, J. J. Postal, PM 1963. “Semantic Interpretation of Idioms and Sentences Containing Them”. Quarterly Progress Report, 70, 275-82.

Katz, J. J. and Fodor, J. A. 1963. “The structure of a semantic theory”. Language, 170-210.

Kay, P. and Fillmore, C. J. 1999. “Grammatical constructions and linguistic generalizations: the *What's X doing Y?* construction”. Language, 1-33.

Kearns, K. 2002. “Light verbs in English”.http://www.ling.canterbury.ac.nz/documents.

Keyser, S. J. and Postal, P. M. 1976. “Beginning English Grammar”. New York: Harper and Row.

Kilgarriff, A. Rychly, P. Smrz, P. and Tugwell, D. 2004. “The sketch engine”. Information Technology, 105, 116.

Kintsch, W. 2001, “Predication”. Cognitive Science, 25 (2), pp. 173–202.

Kiparsky, P. 1976. “Oral poetry: some linguistic and typological considerations”. Oral literature and the formula, 73-106.

Koestler, A. 1967. “The ghost in the machine”. London: Hutchinson.

Konopka, A. E. and Bock, K. 2009. “Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production”. Cognitive Psychology, 58(1), 68-101.

Korhonen, A. 2002. “Subcategorization Acquisition”. Ph.D. thesis, University of Cambridge.

Labov, W. 1978. “Denotational structure”. In: Papers from the Parasession on the Lexicon, 220-260. Chicago: Chicago Linguistic Society

Lakoff, G. 1965. “On the nature of syntactic irregularity”. The Computation Laboratory of Harvard University. Mathematical linguistics and automatic translation. Report No. NSF-16.

Lakoff, G. and Johnson, M. 1980. “Metaphors we live by”. University of Chicago Press

Landauer, T. K. and Dumais, S. T. 1997. “A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge”. Psychological review, 104(2), 211.

Langacker, R. W. 1987. “Foundations of cognitive grammar: Theoretical prerequisites (Vol. 1)”. Stanford University Press.

Langacker, R. W. 2000. “A dynamic usage-based model”. Usage-based models of language, 1-63.

Lasnik, H. and Stowell, T. 1991. “Weakest crossover”. Linguistic inquiry, 687-720

Lenci, A. 2008. “Distributional semantics in linguistic and cognitive research”. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, Special Issue of the Italian Journal of Linguistics, 20(1), 1-31.

Lenci, A. 2014. “Carving verb classes from corpora”. Word Classes. Nature, typology and representations, Current Issues in Linguistic Theory, 17-36.

Lenci, A. Lapesa, G. and Bonansinga, G. 2012. “LexIt: A Computational Resource on Italian Argument Structure”. In LREC (pp. 3712-3718).

Lenci, A. Lebani, G. E. Castagnoli, S. Masini, F. and Nissim, M. 2014. “SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations”. The First Italian Conference on Computational Linguistics CLiC-it 2014, 234-238

Lenci, A. Lebani, G. E. Senaldi, M. S. G. Castagnoli, S. Masini, F. and Nissim, M. 2015 “Mapping the Constructicon with SYMPAThy. Italian Word Combinations between fixedness and productivity”. NetWordS 2015 Word Knowledge and Word Usage, 144.

Levelt, W. J. M. 1989. “Speaking: From intention to articulation”. Cambridge, MA: MIT Press.

Levin, B. and Rappaport Hovav, M. 2005. “Argument realization”. Cambridge University Press.

Li, L. and Sporleder, C. 2009. “A cohesion graph based approach for unsupervised recognition of literal and non-literal use of multiword expressions”. In Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (pp. 75-83). Association for Computational Linguistics.

Li, L. and Sporleder, C. 2009. “Classifier combination for contextual idiom detection without labelled data”. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 315-323). Association for Computational Linguistics.

Li, L. and Sporleder, C. 2010a. “Linguistic cues for distinguishing literal and non-literal usages”. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 683-691). Association for Computational Linguistics

Li, L. and Sporleder, C. 2010b. “Using Gaussian Mixture Models to detect figurative language in context”. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 297-300). Association for Computational Linguistics.

Li, L. B. Roth, and C. Sporleder, 2010, “Topic models for word sense disambiguation and token-based idiom detection,” in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1138–1147.

Li, L. Roth, B. and Sporleder, C. 2010. “Topic models for word sense disambiguation and token-based idiom detection”. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1138-1147). Association for Computational Linguistics

Libben, M. R. and Titone, D. A. 2008. “The multidetermined nature of idiom processing”. Memory and Cognition, 36(6), 1103-1121.

Lieven, E. Behrens, H. Speares, J. and Tomasello, M. 2003. “Early syntactic creativity: A usage-based approach”. Journal of child language, 30(02), 333-370

Light, M. and Greiff, W. 2002. “Statistical models for the induction and use of selectional preferences”. Cognitive Science, 87:1–13.

Lin, D. 1999, “Automatic identification of non-compositional phrases,” in Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL99) p. 317-324.

Lin, D., & Pantel, P. 2001. “Discovery of inference rules for question-answering”. Natural Language Engineering, 7(04), 343-360

Link, G. 1983. “The Logical Analysis of Plurals and Mass Terms”, in Bauerle, R. Schwarze, C. and von Stechow, A. (eds.), Meaning, Use, and Interpretation of Language, pp. 302-323.

Longobardi, G. 1992. “Proper Names and the Theory of N-movement in Syntax and

Logical Form". Working Papers in Linguistics, 1, 1991

Longobardi, G. 2001. "How comparative is semantics? A unified parametric theory of bare nouns and proper names". Natural Language Semantics, 9, 335–369.

Losiewicz, B. L. 1992. "The effect of frequency on linguistic morphology". University of Texas.

Lounsbury, F. G. 1963. "Linguistics and psychology". In S. Koch (Ed.), Psychology: Study of a science (pp. 553–582). New York: McGraw–Hill.

Lyons, J. 1968. "Introduction to theoretical linguistics". Cambridge University Press.

Maher, Z. 2013. "Opening a Can of Worms: Idiom Flexibility, Decomposability, and the Mental Lexicon", PhD Thesis, Yale University.

Makkai, A. 1972. "Idiom Structure in English". The Hague: Mouton

Manning, C. D. and Schütze, H. 1999. "Foundations of statistical natural language processing". MIT press.

Marshall, J. "Syntactic analysis as a part of understanding". Bulletin of the British Psychological Society, 1965,18,28.

Masini, F. 2009, "Combinazioni di parole e parole sintagmatiche" in: Maria Catricalà, Paola Pietrandrea, Edoardo Lombardi Vallauri, Paolo Di Giovine, Donato Cerbasi, Lunella Mereu, Livio Gaeta, Giuliana Fiorentino, Paolo D'Achille, Maria Grossmann, Elisabetta Jezek, Francesca Masini, Anna Pompei, Elisabetta Bonvino, Franca Orletti, Mara Frascarelli. "Spazi linguistici. Studi in onore di Raffaele Simone", ROMA, Bulzoni, pp. 191 – 209

Masini, F. 2012. "Parole sintagmatiche in italiano", Roma, Caissa

Matthews, D. and C. Bannard. 2010. "Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child directed speech". Cognitive Science, 34 (3), 465–488

McCarthy, D. B. Keller, and J. Carroll, 2003, "Detecting a continuum of compositionality in phrasal verbs," in Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, p. 73-80.

McGlone, S. M. Glucksberg, S. and Cacciari, C. 1994. "Semantic productivity and idiom comprehension". Discourse Processes, 17, 167–190.

McRae, K. Ferretti, T. Amyote, L. "Thematic roles as verb-specific concepts". Language and Cognitive Processes: Special Issue on Lexical Representations in Sentence Processing, 12 (1997), pp. 137–176.

Medin, D. L. and Schaffer, M. M. 1978. "Context theory of classification learning". Psychological review, 85(3), 207.

Mel'čuk, I. 1995. "Phrasemes in language and phraseology in linguistics", in Everaert, M. Van Der Linden, E. Schenk, A. and Schreuder, R. (eds.), "Idioms: Structural and Psychological Perspectives", Hillsdale, Lawrence Erlbaum, 167-232.

Miller, G. A. and Johnson-Laird, P. N. 1976. “Language and perception”. Belknap Press

Mitchell, J. and M. Lapata, 2010, “Composition in distributional models of semantics,” Cognitive Science, vol. 34, no. 8, pp. 1388–1429.

Moon, R. 1998. “Fixed expressions and idioms in English: A corpus-based approach”. Oxford University Press.

Muzny, G. and Zettlemoyer, L. S. 2013. “Automatic Idiom Identification in Wiktionary”. In EMNLP (pp. 1417-1421).

Nattinger, J. R. and De Carrico, J. 1994. “Lexical phrases for communicative language teaching”. Paper presented at 28th TESOL Convention, Baltimore, MD.

Nattinger, J. R. and DeCarrico, J. S. 1992, “Lexical Phrases and Language Teaching”, Oxford, Oxford University Press.

Newmeyer, F. J. 1974. “The regularity of idiom behavior”. Lingua, 34(4), 327-342.

Nissim, M. and Zaninello, A. 2013. “Modeling the internal variability of multiword expressions through a pattern-based method”. ACM Trans. Speech Lang. Process. 10(2):1–26.

Nissim, M. Castagnoli, S. and Masini, F. 2014. “Extracting MWEs from Italian corpora: A case study for refining the POS-pattern methodology”. In Proceedings of the 10th Workshop on Multiword Expressions (MWE), pages 57–61.

Nosofsky, R. M. 1988. “Similarity, frequency, and category representations”. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(1), 54.

Nunberg, G. 1978. “The pragmatics of reference”. Indiana University Linguistics Club.

Nunberg, G. I. Sag, and T. Wasow, 1994, “Idioms,” Language, vol. 70, no. 3, pp. 491–538.

Odijk, J. 2004. “A proposed standard for the lexical representation of idioms”. In Proceedings of EURALEX. 153–164.

Ortony, A. Schallert, D. L. Reynolds, R. E. and Antos, S. J. 1978. “Interpreting metaphors and idioms: Some effects of context on comprehension”. Journal of verbal learning and verbal behavior, 17(4), 465-477.

Oxford English Dictionary. 1989. Oxford: Oxford University Press.

Padó, S. and Lapata, M. 2007, “Dependency-based construction of semantic space models”, Computational Linguistics, 33/2, pp. 161-199

Paul, H. 1920[1880], “Prinzipien der Sprachgeschichte”, Halle, Niemeyer.

Pawley, A. and Syder, F. H. 1983. “Two puzzles for linguistic theory: Nativelike selection and nativelike fluency”. Language and communication, 191, 225.

Pearce, D. 2001. “Synonymy in collocation extraction”. In *NAACL 2001 Workshop: WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Carnegie Mellon University, Pittsburgh, June.

Pearce, D. 2002. “A Comparative Evaluation of Collocation Extraction Techniques”. In LREC 2002, Proceedings

Peng, J. A. Feldman, and E. Vylomova, 2014, “Classifying idiomatic and literal expressions using topic models and intensity of emotions,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 2019–2027.

Perek, F. 2014. “Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony”. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland USA, June 23-25 2014.

Perek, F. 2015. “Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives”. Amsterdam: John Benjamins

Perek, F. to appear. “Using distributional semantics to study syntactic productivity in diachrony: A case study”

Peterson, R. R. Burgess, C. Dell, G. S. and Eberhard, K. M. 2001. “Dissociation between syntactic and semantic processing during idiom comprehension”. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(5), 1223

Pianta, E. Bentivogli, L. and Girardi, C. 2002. “MultiWordNet: developing an aligned multilingual database”. In Proceedings of the first international conference on global WordNet (Vol. 152, pp. 55-63).

Pinker, S. 1989. “Learnability and cognition: The acquisition of argument structure”. Cambridge, MA: MIT Press/Bradford Books

Pinker, S. 1995. “The language instinct: The new science of language and mind” (Vol. 7529). Penguin UK.

Pinker, S. and Jackendoff, R. 2005. “The faculty of language: what's special about it?”. Cognition, 95(2), 201-236.

Piunno, V. Masini, F. and Castagnoli, S. 2013. “Studio comparativo dei dizionari combinatori dell’italiano e di altre lingue europee”. CombiNet Technical Report. Roma Tre University and University of Bologna

Popiel, S. J. and McRae, K. 1988. “The figurative and literal senses of idioms, or, all idioms are not used equally”. Journal of Psycholinguistic Research, 17, 475–487

Posner, M. I. and Keele, S. W. 1968. “On the genesis of abstract ideas”. Journal of experimental psychology, 77(3p1), 353.

Quillian, R. “Semantic memory”. In M. Minsky (Ed.), “Semantic information processing”. Cambridge: M.LT. Press. 1968.

R Core Team, 2015, “R: A Language and Environment for Statistical Computing”, R Foundation for Statistical Computing, Vienna

Ramisch, C. Schreiner, P. Idiart, M. and Villavicencio, A. 2008. “An Evaluation of

Methods for the Extraction of Multiword Expressions". In Proceedings of the LREC Workshop MWE pages 50–53.

Ramisch, C. Schreiner, P. Idiart, M. and Villavicencio, A. 2008. "An evaluation of methods for the extraction of multiword expressions". In Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008) (pp. 50-53).

Rappaport Hovav, M. and Levin, B. 1998. "Building verb meanings". The projection of arguments: Lexical and compositional factors, 97-134.

Redfern, W. D. 1989. "Clichés and coinages". Basil Blackwell.

Resnik, P. S. 1993. "Selection and information: a class-based approach to lexical relationships". IRCS Technical Reports Series, 200.

Riehemann, S. 2001. "A Constructional Approach to Idioms and Word Formation". Ph.D. thesis, Stanford University.

Ritz, J. and U. Heid, 2006, "Extraction tools for collocations and their morphosyntactic specificities," in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC06), Genoa, p. 1925-1930.

Rizzi, L. 1990. "Relativized Minimality". The MIT Press.

S. Schulte im Walde. 2008. "The induction of verb frames and verb classes from corpora". In Lüdeling A. and Kytö M. editors, Corpus Linguistics: An International Handbook, chapter 61. Mouton de Gruyter, Berlin

Sag, I. A. 1976. "Deletion and logical form" (Doctoral dissertation, Massachusetts Institute of Technology).

Sag, I. A. T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, 2001, "Multiword expressions: A pain in the neck for NLP" in Proc. Of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), pp. 1–15.

Sahlgren, M. 2006, "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces", PhD Thesis

Sahlgren, M. 2008, "The Distributional Hypothesis". Italian Journal of Linguistics, 20 (1), pp. 33-53.

Salton, G., Wong, A. and Yang, C.-S. 1975. "A vector space model for automatic indexing". Communications of the ACM, 18 (11), pp. 613-620.

Saussure, F. D. 1966[1916]. "Course de Linguistique Générale", translation by Wade Baskin, New York: McGraw-Hill Book Co

Schone, P. and Jurafsky, D. 2001. "Is knowledge-free induction of multiword unit dictionary headwords a solved problem?". In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (pp. 100-108).

Searle, J. 1975. "Indirect speech acts". In Cole, P. and Morgan, J. L. (eds.), "Syntax and Semantics. Speech Acts". New York: Academic Press, 59–82.

Seretan, V. 2011. “Syntax-based collocation extraction (Vol. 44)”. Springer Science and Business Media.

Seretan, V. Nerima, L. and Wehrli, E. 2003. “Extraction of multi-word collocations using syntactic bigram composition”. In Proceedings of RANLP- 03, pages 424–431.

Seretan, V. Nerima, L. and Wehrli, E. 2003. “Extraction of multi-word collocations using syntactic bigram composition”. Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003). Borovets (Bulgaria) p. 424-431

Shannon, C.E. 1948. “A mathematical theory of communication”. The Bell System Technical Journal, 27(3):379 – 423.

Simone, R. 1990, “Fondamenti di linguistica”, Roma - Bari, Laterza.

Sinclair, J. 1991. “Corpus, concordance, collocation”. Oxford University Press.

Siyanova-Chanturia, A. and R. Martinez, 2014, “The idiom principle revisited”, Applied Linguistics, pp. 1–22

Smadja, F. 1993. “Retrieving collocations from text: Xtract”. Computational linguistics, 19(1), 143-177.

Sorhus, H. B. 1977. “To Hear Ourselves--Implications for Teaching English as a Second Language”. English Language Teaching Journal, 31(3), 211-221.

Sporleder, C. and Li, L. 2009. “Unsupervised recognition of literal and non-literal use of idiomatic expressions”. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 754-762). Association for Computational Linguistics.

Sprenger, S. A. Levelt, W. J. and Kempen, G. 2006. “Lexical access during the production of idiomatic phrases”. Journal of memory and language, 54(2), 161-184

Sprouse, J. 2011. “A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory”. Behavior research methods, 43(1), 155-167.

Squillante, L. 2014. “Towards an empirical subcategorization of multiword expressions”. In Proceedings of the 10th Workshop on Multiword Expressions (MWE), pages 77–81, Gothenburg, Sweden, April. Association for Computational Linguistics.

Squillante, L. 2014. “Towards an empirical subcategorization of multiword expressions”. EACL 2014, 77.

Strässler, J. 1982. “Idioms in English: A pragmatic analysis” (Vol. 183). Gunter Narr Verlag.

Stroop, J. R. 1935. “Studies of interference in serial verbal reactions”. Journal of experimental psychology, 18(6), 643.

Swinney, D. A. and Cutler, A. 1979. “The access and processing of idiomatic expressions”. Journal of verbal learning and verbal behavior, 18(5), 523-534.

Tabossi, P. Fanari, R. and Wolf, K. 2008. “Processing idiomatic expressions: Effects of

semantic compositionality". Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(2), 313.

Tabossi, P. L. Arduino, and R. Fanari, 2011, "Descriptive norms for 245 italian idiomatic expressions," Behavior Research Methods, vol. 43, pp. 110–123

Tabossi, P. Wolf, K. and Koterle, S. 2009. "Idiom syntax: Idiosyncratic or principled?". Journal of Memory and Language, 61(1), 77-96.

Tannen, D. 1989. "Talking Voices: Repetition. Dialogue, and Imagery in Conversational Discourse", Cambridge, New York.

Tapanainen, P. Piitulaine, J. and Jarvinen, T. 1998. "Idiomatic object usage and support verbs". In 36th Annual Meeting of the Association for Computational Linguistics.

Tiberii, P. 2012. "Dizionario delle collocazioni. Le combinazioni delle parole in italiano". Zanichelli

Titone, Debra A. and Connine, C. M. 1994. "Descriptive norms for 171 idiomatic expressions: familiarity, compositionality, predictability, and literality". Metaphor and Symbolic Activity 9(4): 247–70

Tomasello, M. 2003. "Constructing a Language: A Usage-Based Theory of Language Acquisition". Harvard University Press.

Tremblay, A. and Baayen, R. H. 2010. "Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall". Perspectives on formulaic language: Acquisition and communication, 151-173.

Tremblay, A. Derwing, B. Libben, G. and Westbury, C. 2011. "Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks". Language Learning, 61(2), 569-613

Turner, N. E. and Katz, A. N. 1997. "The availability of conventional and of literal meaning during the comprehension of proverbs". Pragmatics and Cognition, 5(2), 199-233.

Turney, P. 2008, "A uniform approach to analogies, synonyms, antonyms and associations", in Proceedings of COLING 2008, pp. 905–912

Turney, P.D. and P. Pantel, 2010, "From Frequency to Meaning: Vector Space Models of Semantics," Journal of Artificial Intelligence Research, vol. 37, pp. 141–188.

Van Gestel, F. 1995. "En bloc insertion". In Everaert, M. van der Linden, E-J. Schenk, A. and Schroeder, R. (eds.) "Idioms: structural and psychological perspectives", 75-94.

Van Lancker, D. and Cummings, J. L. 1999. "Expletives: neurolinguistic and neurobehavioral perspectives on swearing". Brain research reviews, 31(1), 83-104.

Van Lancker-Sidtis, D. and Rallon, G. 2004. "Tracking the incidence of formulaic expressions in everyday speech: Methods for classification and verification". Language and Communication, 24(3), 207-240.

Van Valin Jr, R. D. and Randy, J. La Polla. 1997. "Syntax", Structure, Meaning and

Function. Cambridge: Cambridge UP.

Venkatapathy, S. and A. Joshi, 2005, “Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features,” in Proceedings of Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP05), Vancouver, pp. 899–906.

Vietri, S. 2014. “Idiomatic Constructions in Italian: A Lexicon-grammar Approach”. John Benjamins Publishing Company

Villavicencio, A. Kordoni, V. Zhang, Y. Idiart, M. and Ramisch, C. 2007. “Validation and evaluation of automatically acquired multiword expressions for grammar engineering”. In Proceedings of EMNLP-CoNLL 2007, pages 1034–1043.

Villavicencio, A. Kordoni, V. Zhang, Y. Idiart, M. and Ramisch, C. 2007. “Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering”. In EMNLP-CoNLL (pp. 1034-1043).

Voghera, M. 2004. “Polirematiche”. Linguistica Pragensia, 67(2):100–108.

Wasow, T. Sag, I. Nunberg, G. 1984. “Idioms: An interim report”. In: Hattori, S. Inoue, K. (Eds.), Proceedings of the XIIIth International Congress of Linguistics, Tokyo.

Watkins, C. 1992. “The comparison of formulaic sequences. Reconstructing Languages and Cultures”, The Hague, 391-418.

Weinreich, U. 1969. “Problems in the analysis of idioms. Substance and structure of language”, 23-81.

Wermter, J. and Hahn, U. 2006. “You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction”. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 785-792). Association for Computational Linguistics.

Weydt, H. 1972. “Unendlicher Gebrauch Von Endlichen Mitteln: Mißverständnisse um ein linguistisches Theorem”. Poetica, 249-267.

Whittlesea, B. W. 1987. “Preservation of specific experiences in the representation of general knowledge”. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(1), 3.

Widdows, D. and Dorow, B. 2005. “Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns”. In Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition (pp. 48-56). Association for Computational Linguistics.

Williams, L. Bannister, C. Arribas–Ayllon, M. Preece, A. and Spasić, I. 2015. “The role of idioms in sentiment analysis”. Expert Systems with Applications.

Wray, A. 2002. “Formulaic Language and the Lexicon”. Cambridge: Cambridge University Press.

Wray, A. 1992. “The focusing hypothesis: the theory of left hemisphere lateralized language re-examined”. Amsterdam: John Benjamins.

Wulff, S. 2008, “Rethinking Idiomaticity: A Usage-based Approach”. Continuum.

Wulff, S. 2009, “Converging evidence from corpus and experimental data to capture idiomaticity,” Corpus Linguistics and Linguistic Theory, vol. 5, no. 1, pp. 131–159.

Zeldes, A. 2013. “Productive argument selection: Is lexical semantics enough?” Corpus Linguistics and Linguistic Theory, 9(2):263–291

Zgusta, L. 1967. “Multiword lexical units”. Word, 23(1-3), 578-587.