# When GL meets the corpus: a data-driven investigation of semantic types and coercion phenomena

**Elisabetta JEZEK**
University of Pavia, Dept. of Linguistics
Strada Nuova 65
Pavia, Italy, 27100
jezek@unipv.it

**Alessandro LENCI**
University of Pisa, Dept. of Linguistics
Via Santa Maria 36
Pisa, Italy, 56126
alessandro.lenci@ilc.cnr.it

**Abstract**

In this paper we present an analysis of corpus-derived V-Object combinations aiming to provide a data-driven characterization of Semantic Types (STs) and improve our understanding of how types behave compositionally, i.e. how they enter compositional processes and are modulated by them. As a theoretical framework, we adopt the enriched compositional rules and the type system as presented in Pustejovsky (2007). Our main concerns are twofold: i.) first of all, we will show with a specific case-study how a data-driven investigation can shed light on the organization of the type system and on semantic compositional operations affecting types; ii.) starting from the results of this investigation, we intend to propose a general methodology for lexical modeling in which the Generative Lexicon (GL) theory and corpus analysis are deeply interwoven in a process of mutual feeding. In fact, we argue that, if on the one hand corpus data can help to anchor the study of lexical dynamics and type system on empirical evidence, on the other hand GL can provide the crucial interpretative key for corpus data.

## 1 Theoretical background

One of the major developments of the GL theory in recent years has been the integration of the type system into a theory of argument selection where what counts for compositional rules is the correspondence between the type selected by the predicate and the type of the argument(s) (Pustejovsky 2001, 2007). Types may be of three main sorts: *simple*-, *unified*- and *dot-types*. Simple types correspond to natural types, e.g. *lion*, *rock*, *water*, etc. Unified types extend simple types with telic and/or agentive dimensions, and essentially correspond to types of artifactual entities and/or entities inherently endowed with a specific functionality, e.g. *knife*, *beer*, *teacher*, etc. Finally, dot types correspond to intrinsically polysemous types (e.g. *school*, *book*, etc.), obtained through a complex type-construction operation on natural and unified types. This tripartite type system also applies to verbs and adjectives, which express simple, unified or dot predicative functions depending on the type of the argument they select. What triggers semantic operations such as *coercion* is precisely the syntagmatic clash between *selecting* and *selected* type. When it occurs, this clash may fail completely to assign an interpretation to the combination (as in the case of **the rock died*) or it may give rise to two kinds of *coercion operations*: *exploitation* and *introduction*. In the first case, some component of the lexical meaning is accessed and exploited, whereas in the second case, some new conceptual material is introduced contextually. Globally, the theory now predicts 9 possible domain-preserving operations on types, as reported in Table 1. Next to operations on types, GL syntagmatic processes also include *co-composition* phenomena between V and argument, which license new interpretations of the predicate in context. Since both operations of *typing* and *co-composition* may take place simultaneously on the same syntagmatic sequence, the picture of *what goes on where* in a word combination, as far as the construction of its meaning goes, is not an easy one to reconstruct.

| | **Type selected** | | |
|---|---|---|---|
| **Argument type** | *Simple (natural)* | *Unified (artifactual)* | *Dot (complex)* |
| *Simple (natural)* | Selection | Introduction | Introduction |
| *Unified (artifactual)* | Exploitation | Selection | Introduction |
| *Dot (complex)* | Exploitation | Exploitation | Selection |

Table 1 – Composition operations on types in GL

## 2 Why and how is corpus evidence crucial for a GL-like semantic theory?

Corpora have often been regarded as a precious source of evidence to feed GL-like lexical models. Various corpus-based techniques have been applied to learn qualia structure information from corpora (cf. Bouillon *et al.* 2002; Yamada & Baldwin 2004). Pustejovsky *et al.* (2004) present a strategy to develop a corpus-driven type system through the use of Corpus Pattern Analysis (CPA), an approach to which the present research is explicitly and most directly related. CPA is a semi-automatic bootstrapping process to produce a dictionary of selection contexts for predicates in a language (Hanks & Pustejovsky 2005). Corpus-derived syntagmatic patterns are mapped onto GL as a linguistic model of interpretation, which guides and constrains the induction of word senses from distributional information. In our research we apply the basic ideas of CPA to explore the organization of the type system and its qualia articulation, as well as the compositional operations that act on STs.

Notwithstanding the richness of evidence on word behavior it provides, the use of corpus analysis raises the crucial issue of how to properly map the extracted patterns onto the GL architecture of the lexicon. Let us call $\sigma$ a given predicative complex V-N extracted from a corpus such as <*eat–cake*$_{obj}$> or <*read–book*$_{obj}$>, etc. Each $\sigma$ is a piece of observed evidence of the distribution of lexical items in context. The key epistemological issue is thus the following: *what kind of inferences we can draw from the extracted contexts σ about the type system and the compositional rules?* Given a certain context $\sigma$ that we observe in a corpus, we have to ask ourselves three sorts of related but independent questions: i.) what is the type of N? ii.) what is the type selected by the V? iii.) what is the particular operation that allowed N and V to compose semantically in $\sigma$? Our claim is that these three questions can be answered by investigating the combinatorial distributions of V and N in a corpus. We assume that the combinatorial distribution of a lexical item is determined and constrained by its type and that for this reason it can be taken as an empirical indicator of what the type is. We expect lexical items belonging to same type to show a similar syntagmatic distribution and differences in distribution to be indicators of differences in type (although we will see later that this assumption is sometimes too strong and needs to be restrained). Notice that this strategy differs radically from other approaches that assume that the type of a given lexical item is provided by a fixed, corpus-independent, fully-fledged ontology of semantic types such as for instance WordNet (Fellbaum, 1998). Although we are not against the idea of adopting a predefined ontology of semantic types, we believe this should rather be conceived as a shallow repository of semantic types (much in the style of the *Brandeis Shallow Ontology*, as described in Pustejovsky *et al.* 2006), that represent the starting point for a corpus-based definition of fine-grained STs emerging as *abstractions over the combinatorial patterns of lexical items*. We thus propose that by inspecting a reasonably large amount of syntagmatic contexts extracted form a corpus it is possible to draw a more detailed map of a GL-style lexical type system.

The key point is that any attempt to get at a data-driven characterization of STs can not dispense with a careful analysis of the compositional operations between types, which are responsible for the empirical distribution of V-N pairs we observe in corpora. Given GL architecture, we have to assume that each context pair $\sigma$ has been generated by the combinations of two different factors: i.) the structure of the STs to which V and N in $\sigma$ belong, as well as their position in the overall type system; ii.) the particular semantic operations that have driven the semantic composition of V and N in $\sigma$. If $\sigma$ represents our empirical observational datum, i.) and ii.) are the two *hidden parameters* that we have to discover. As we said above in §. 1, given the assumption that compositionality is not driven by pure type selection only, the challenge for any corpus-based approach to GL is exactly how to reconstruct the complex interplay between the type system and the array of semantic operations on types that we have to assume as being operative in every syntagmatic context.

## 3 Corpus processing and data extraction

In this research we focus our attention on Italian data, although we believe that most of our claims extend to other languages quite straightforwardly. Our dataset includes 877,352 syntagmatic contexts $\sigma$ of V-N pairs, in which N is either the subject (374,948) or the direct object (502,404) of V. In this paper we have focused only on V-obj contexts. Each token $\sigma$ has been automatically extracted from a 20 million subset of the *La Repubblica Corpus,* a 450 million word corpus of written Italian newspaper articles (Baroni *et al.* 2004). The corpus subset has been automatically processed

with IDEAL+ (Bartolini *et al.* 2004), a rule-based, finite-state dependency parser for Italian. From the parser outputs we extracted the context pairs that we used to build *lexical sets* for nouns and verbs. Following Hanks & Pustejovsky (2005), and Hanks (2006), we define the *lexical set* LS for a noun N (or for a verb V) as the list of verbs (nouns) with which the noun (verb) typically occurs as direct object. In other words, LSs are paradigmatic series of words that can occupy the same syntagmatic position (either as argument or predicate). We will see later how this notion is crucial in our investigation. In order to anchor the notion of typical co-occurrence on firmer quantitative grounds, we used *log-likelihood* (Dunning 1993) to measure the strength of association between each V and N type in our dataset. The elements of LS of a noun N with the highest log-likelihood score therefore represent the most typical predicates with which N occurs as direct object: we will refer to such sets as *verbal LSs*. Symmetrically, the elements of LS of a verb V with the highest log-likelihood score are the most typical nouns that occur as direct objects of V; these sets will be referred below as *nominal LSs*. Although we are perfectly aware that our definitions of $\sigma$ and of LS abstract away from many important features of the whole word context (e.g. the presence of other arguments, modifiers, etc.), they nevertheless reveal interesting properties of the lexical type system, as our analysis below will show.

## 4 Anatomy of a type: the case of *leggere* "read"

The rest of this paper is devoted to present a case study in which the methodology illustrated above is applied to an in-depth analysis of the semantic type associated with the verb *leggere* "read". In particular, in this section we aim at showing how the nouns appearing in the LS of this verb can be projected on a GL ontology of semantic types described in terms of their qualia structure, while in §. 5 the same empirical data will provide evidence for a more complex articulation of the lexical type system. In §. 6, corpus analysis will be used to explore the operations proposed in GL to describe the compositional dynamics between predicates and their arguments.

First of all, why *leggere*? The reason of choosing this verb as the starting point for our case study of a specific semantic type is that its English equivalent *read* is a predicate whose selective environment is *prima facie* fairly well-characterized within GL. In fact, it is defined as a complex functional type selecting for a complex, dot-argument as its direct object: $\lambda y{:}phys \bullet info$ $\lambda x{:}e_N$ [*read*(x,y)]. This analysis is motivated by the fact that "the concept of reading is *sui generis* to an entity that is defined as 'informational print matters', that is, a complex type such as *phys* ● *info*" (Pustejovsky 2007: 29). Consequently, given the battery of semantic operations illustrated in §. 1 above, we expect pure selection to apply between *read* and whatever lexical item that is an instance of this dot-type. The prototypical case of this sort of composition occurs in the phrase *read the book*: "the predicate read requires a dot object of type *phys* ● *info* as its direct object, and the NP present, *the book*, satisfies this typing directly" (ibid.: 32).

Lexical sets as defined in §. 3 can be used to carry out a sort of "autoptic analysis" of types in order to evaluate whether our intuition about the selective environment of *leggere* is validated and simultaneously refined with the help of text-driven data. To this purpose, we extracted from our dataset the nominal LS of *leggere*, which includes the most typical nouns occurring as direct object of this predicate in our corpus. In Table 2 we reported the top 40 nouns of this nominal LS, ordered by decreasing log-likelihood (ll) values. If we look at this table, we immediately see that the lexical set of nouns combining with *leggere* does not directly map to a single semantic type, and that from the fact that a noun is included in the nominal lexical set of *leggere,* we can not simply infer that the type of the noun is *phys* ● *info*. The reason for this is twofold, and is consistent with GL predictions: first of all, *leggere* has the ability not only to combine by pure selection, but also to coerce the argument type. This is the case for instance of person names like *Freud* and *Rimbaud* occurring in the nominal LS of *leggere*, and that are clearly coerced to be interpreted as the works written by these authors. Secondly, *leggere* can itself undergo co-compositions when combining with an argument that does not match its selective requirements and licence different meanings, as in the case of *leggere il pensiero*, where *leggere* = 'interpret'

Taking this into account, it becomes cleat that the analysis of LS brings afore a truly general methodological issue, i.e. *what does the fact of observing a given noun within the lexical set of a verb tell us about the noun's type as well as its internal structure*? We would like to claim that this problem can be dealt with by reversing the perspective of the analysis and inspecting the composition of the verbal LSs of the nouns, looking at two aspects simultaneously: the selectional properties of the verbs, and their association strength (ll value). This actually means that we have to explore a larger area of the combinatorial space of lexical items: i.e. we can try

| noun | ll value | noun | ll value | noun | ll value |
|---|---|---|---|---|---|
| *libro* "book" | 225,44 | *cartella* "page" | 40,64 | *missiva* "missive" | 15,85 |
| *giornale* "newspaper" | 174,98 | *messaggio* "message" | 36,10 | *telegramma* "telegram" | 14,97 |
| *articolo* "article" | 133,28 | *relazione* "report" | 35,14 | *poesia* "poem" | 14,77 |
| *lettera* "letter" | 96,77 | *passo* "passage" | 34,60 | *verdetto* "verdict" | 14,62 |
| *romanzo* "novel" | 76,63 | *resoconto* "report" | 30,04 | *brano* "passage" | 14,62 |
| *testo* "text" | 58,34 | *parola* "word" | 29,71 | *nota* "note" | 14,51 |
| *documento* "document" | 56,42 | *frase* "sentence" | 28,75 | *opera* "work" | 14,20 |
| *intervista* "interview" | 52,37 | *sentenza* "sentence" | 25,93 | *Rimbaud* | 14,19 |
| *comunicato* "communiqué" | 49,23 | *motivazione* "reason" | 23,39 | *sofisma* "sophisma" | 14,19 |
| *dichiarazione* "statement" | 48,07 | *Freud* | 19,96 | *Tuttosport* | 14,19 |
| *pagina* "page" | 47,76 | *Financial Times* | 19,40 | *scritta* "writing, notice" | 11,75 |
| *sceneggiatura* "script" | 44,17 | *omelia* "sermon" | 16,92 | *telex* "telex" | 11,59 |
| *riga* "line" | 42,03 | *notizia* "news" | 16,14 | | |
| *discorso* "speech" | 41,07 | *saggio* "essay" | 16,04 | | |

Table 2 - top 40 nouns in the LS of *leggere*

| libro "book" | articolo "article" | testo "text" |
|---|---|---|
| *scrivere* "write" | *scrivere* "write" | *pubblicare* "publish" |
| *leggere* "read" | *leggere* "read" | *approvare* "approve" |
| *pubblicare* "publish" | *pubblicare* "publish" | *votare* "vote" |
| *presentare* "present" | *inviare* "send" | *leggere* "read" |
| *sfogliare* "leaf through" | *ricevere* "receive" | *modificare* "modify" |
| *dedicare* "dedicate" | *abrogare* "cancel" | *scrivere* "write" |
| *riscrivere* "rewrite" | *applicare* "enforce" | *redigere* "write" |
| *tradurre* "traslate" | *dedicare* "dedicate" | *emendare* "amend" |
| *ristampare* "reprint" | *approvare* "approve" | *preparare* "prepare" |
| *vendere* "sell" | *bocciare* "reject" | *diffondere* "circulate" |
| **romanzo "novel"** | **lettera "letter"** | **messaggio "message"** |
| *scrivere "write"* | *inviare "send"* | *inviare "send"* |
| *leggere "read"* | *scrivere "write"* | *lanciare "send"* |
| *pubblicare "publish"* | *ricevere "receive"* | *mandare "send"* |
| *ristampare "reprint"* | *spedire "send"* | *ricevere "receive"* |
| *concepire "conceive"* | *leggere "read"* | *consegnare "deliver"* |
| *intitolare "give a title"* | *mandare "send"* | *trasmettere "transmit"* |
| *pianificare "plan"* | *recapitare "deliver"* | *intercettare "intercept"* |
| *filmare "film"* | *consegnare "deliver"* | *leggere "read"* |
| *comprare "buy"* | *pubblicare "publish"* | *portare "bring"* |
| *finire "finish"* | *firmare "sign"* | *recapitare "deliver"* |

Table 3 - top 10 verbs in the LS of a set of nouns in the LS of *leggere*

to gain some insights about the selecting type of a predicate V by looking at the other verbs $\{V_{ij},\ldots,V_{kj}\}$ with which a noun $N_j$ combines, with $N_j$ a member of the nominal LS of V. Notice, however, that this operation is not straightforward for the same reason we mentioned for *leggere*. Verbal LSs may contain two sorts of verb: *best verbs*, i.e. verbs that match the noun type and combine by pure selection, and *coercing verbs*, i.e. verbs that do not match the noun type and coerce it either via exploitation or introduction. Within the most frequent $\sigma$, we can thus expect to find both these verbs, although in principle we assume introductions to be more likely situated in low frequencies of $\sigma$.

Keeping this in mind, we have extracted the verbal LS of a subset of 6 nouns co-occurring with *leggere* in Table 2. These nouns are: *libro* "book", *articolo* "article", *testo* "text", *romanzo* "novel", *lettera* "letter", *messaggio* "message". For reasons of space, we have reported in Table 3 only the top 10 verbs (ordered for decreasing ll values) of the verbal LSs of these nouns. The analysis of these LSs bring afore interesting regularities and enables us to identify two first subsets of nouns, which we discuss below:

- *libro* "book", *articolo* "article", *testo* "text", *romanzo* "novel". The verbal LSs of these nouns all share the fact of being characterized by verbs expressing acts of composing or using semiotic artifacts in which the printed dimension is at least as salient as the informational one. In fact, in the top ranks of these LSs we find verbs expressing variations of writing (e.g. *scrivere*, *riscrivere*, etc.), reading (*leggere*, *rileggere*, *leggiucchiare*, etc.) and printing (e.g. *pubblicare*, *stampare*, *ristampare*, etc.);

- *lettera* “letter” and *messaggio* “message”. This set is also characterized by verbal LSs dominated by verbs selecting the physical and the informational dimensions. However, now the physical dimension is not selected by events of writing or printing, but rather by events of transmission and exchange (e.g. *mandare*, *inviare*, *spedire*, *ricevere*, etc.).

From this first piece of analysis, we can conclude that there are reasons to believe that these nouns all belong to the type *phys* ● *info*, since they all typically co-occur with verbs selecting for *phys* ● *info* or, alternatively, with verbs selecting for the physical dimension (*portare, posare*) or the informational one (*criticare, censurare, votare)*. However, the question arises how we can account for the differences in their LSs. It is evident that types are not sufficient to account for the whole syntagmatic distribution of these nouns: they do not capture all facets of the semantic of these lexical items. We claim that GL model can provide the right interpretive key for such distributional facts and that the differences in the lexical sets of these nouns can be accounted for in terms of differences in their qualia specifications. Therefore, we believe that the following type representation would be appropriate for the two subsets of nouns discussed above (using the notation of tensor types in Pustejovsky 2007):

(1) *libro* “book”, *articolo* “article”, *romanzo* “novel”, testo “text”:

*phys* ● *info* $\otimes_{\text{Telic}}$ READING_EVENTS {*read*, *reread*,…} $\otimes_{\text{Agentive}}$ WRITING_EVENTS {*write*, *rewrite*, …} $\otimes_{\text{Agentive}}$ PUBLISHING_EVENTS {*publish*, *print*, …}

(2) *lettera* “letter”, *messaggio* “message”:

*phys* ● info $\otimes_{\text{Telic}}$ READING_EVENTS {*read*, *reread*,…} $\otimes_{\text{Telic}}$ TRANSMISSION_EVENTS {*send*, *circulate*, *deliver*…} $\otimes_{\text{Agentive}}$ WRITING_EVENTS {*write*, *modify*, …} $\otimes_{\text{Agentive}}$ PUBLISHING_EVENTS {*publish*, …}

The representations in (1) and (2) also closely correspond to most natural intuitions about the semantics of a noun like *letter*: a *letter*, like a *book* is an artifact created with the purpose of being read. However, the former also differs from the latter because a letter has a further telic dimension concerning transmission: something is not a letter, unless it is designed in such a way that it can be sent or exchanged. Besides, nouns such as *articolo* and *testo* also exhibit in their verbal LS a number of verbs expressive events of the legislative domain (e.g. *approvare*, *votare*, etc.): in fact within the realm of written semiotic artifacts we should account for those endowed with normative and performative character. It is worth emphasizing that these data call for much more advanced models of the type system than those simply couched in terms of taxonomic structures and the like. In this respect, a system like GL, in which fine-grained distinctions can be captured by the way qualia information enters into the type constitution, is able to offer more promising accounts of noun (and verb) semantic properties as emerging from their distributional behaviour.

## 5 Discovering lexical types

Besides providing a refined representation of the nouns as far as their qualia structure is concerned (§. 4), the investigation of the verbal LSs also allows us to confirm empirically our assumptions that the nouns of the verbal LS of *leggere* do not all belong to the same type. Consider again the nouns discussed in the previous sections and compare them to the verbal LSs of *giornale* “newspaper” on the one side, and to *intervista* “interview”, *discorso* “speech”, *dichiarazione* “declaration” reported in Table 4. Although all the nouns in this latter group share *leggere* as one of their most frequent co-occurring verbs, the composition of their verbal LSs differs radically from the ones of the nouns in Table 3.

If we look at the verbal LS of *giornale*, the presence of verbs that typically select for humans or organizations - like *querelare* “bring an action against”, *dirigere* “edit”, *attaccare* “attack” and *obbligare* “force” clearly bring afore an additional key aspect of the polysemy of this noun, i.e. its organizational dimension, that is not at all shared by the lexemes discussed in §. 4. This confirms and at the same time supports our intuition that *giornale* is actually part of a more complex dot type than *phys* ● *info*, i.e. *organization* ● (*phys* ● *info*), and that its representation should therefore be the following:

(3) *giornale* “newspaper”:

*organization* ● (*phys* ● *info* $\otimes_{\text{Telic}}$ READING_EVENTS {*read, ...*} $\otimes_{\text{Agentive}}$ PUBLISHING_EVENTS {*publish, print, ...*}) $\otimes_{\text{Telic}}$ AGENTIVE_EVENTS {*edit, attack, ...*}

Let us now look at the verbal LS of *intervista* “interview”, *discorso* “ speech”, and *dichiarazione* “declaration” in Table 4. What immediately comes into sight is that the physical and/or printed dimension is now in the background: although these nouns co-occur with verbs selecting for physical objects and informational content, they very often combine with verbs that select for the oral/sound dimension (e.g. *pronunciare*, *ascoltare*, *registrare*, etc.) or for the eventive, time enduring

| giornale "newspaper" | intevista "interview" | dichiarazione "declaration" | discorso "speech" |
|---|---|---|---|
| *leggere* "read" | *rilasciare* "give" | *rilasciare* "make" | *pronunciare* "pronounce" |
| *scrivere* "write" | *concedere* "give" | *fare* "make" | *riprendere* "continue" |
| *stampare* "print" | *leggere* "read" | *diffondere* "circulate" | *fare* "make" |
| *sfogliare* "leaf through | *dare* "give" | *leggere* "read" | *tenere* "give" |
| *leggiucchiare* "read" | *mandare* "send" | *presentare* "present" | *leggere* "read" |
| *querelare* "bring an action" | *pubblicare* "publish" | *firmare* "sign" | *allargare* "enlarge |
| *rileggere* "re-read" | *rileggere* "reread" | *sottoscrivere* "endorse" | *pronunziare* "pronounce" |
| *attaccare* "attack" | *realizzare* "make" | *smentire* "refute" | *ascoltare* "listen" |
| *dirigere* "edit" | *raccogliere* "collect" | *consegnare* "deliver" | *rivolgere* "address" |
| *riempire* "fill" | *registrare* "record" | *interpretare* "interpret" | *concludere* "conclude" |

Table 4 - top 10 verbs in the LS of a set of nouns of the LS of *leggere*

character of the entities to which the nouns refer to (e.g. to event-selecting verbs like *concludere*, *riprendere*). Most notably, light verbs (*dare, fare, tenere* etc.), i.e. verbs that typically combine with nouns denoting events, also occupy a central position in the verbal LSs of these nouns.

We claim that the reason why it is so is that these nouns are in fact first of all events with certain temporal duration in which an amount of information is exchanged, primarily orally. This does not imply that interviews, speeches and declarations can not be written or read, but that this dimensions might not be part of their intrinsic denotation. Rather, we would claim that with these nouns the written, physical dimension is coerced, or better introduced to them, by specific verbs, such as *write* or *read*, that can occur with them, and that the type associated to these nouns is *event* ● *info*. As in §. 4, we can express the semantic properties of these nouns with the following type representation (using the notation of tensor types in Pustejovsky 2007):

(4) discorso "speech", intervista "interview" *dichiarazione* "declaration":

*event* ● *info* $\otimes_{\text{Agentive}}$ SPEECH_EVENTS {*pronounce*, *address*, *give a speech*…} $\otimes_{\text{Telic}}$ LISTENING_EVENTS {*listen*, …}

To sum up, from the analysis of the verbal LSs carried out in §. 4 and 5, we may conclude that the variations in the verbal LSs can be interpreted as an indicator of two main facts: *differences in qualia specifications* or *difference in type*. Although some exceptions can de detected, and although we are perfectly aware that our analysis above greatly underestimates the complexity of the lexical type space, our investigation so far shows that the assumptions about what the type of a noun is are sensibly confirmed by and reflected in its syntagmatic behaviour, and that the method of combinatorial analysis of LSs that we have sketched here offers a promising perspective to integrate type system investigation with corpus analysis.

## 6 An overall map of compositional operations

Besides allowing us to confirm or falsify our hypotheses about what the semantic type associated to specific nouns is, corpus analysis can help us to improve our understanding of how types behave compositionally, and thus to contribute to represent how the meaning of a V-N combination is computed. As we already clarified, our starting assumption is that a key property of types is their ability to undergo modifications (coercions) in context, thus expanding exponentially the creative ways in which we can use them to express meanings. Also, following Pustejovsky (2007), we assume that predicates activate coercions on types if these latter do not correspond to the selectional restrictions. We would like to claim that it is precisely these assumptions that corpus analysis can help us to verify, possibly giving us new insights on how we can approach these problems.

Taking Table 1 as the skeleton of our analysis, we see that the GL organization of the type system makes two specific predictions concerning the compositional modes of dot-types, with respect to domain preserving operations: i.) a dot-argument will compose either by pure selection, with a dot-predicate, or by exploitation, with a natural or artifactual selecting predicates (third row of Table 1); ii.) a dot-selecting predicate will compose either by pure selection, with a matching dot-argument, or by introduction, with natural and artifactual arguments (third column of Table 1). Corpus data can be used to verify to what extent these predictions are borne out.

To test the first prediction, we use the verbal LSs of the nouns discussed above, that as a result of our analysis in §. 4 and 5 have been assigned either to the *phys* ● *info* type (e.g. *libro*, *romanzo*, *articolo*, *testo, lettera, messaggio*) or to the *event* ● *info* type (e.g. *intervista*, *discorso*, *dichiarazione*), or to the *organization* ● (*phys* ● *info*) type (i.e. *giornale*). These LSs show that prediction i.) is substantially confirmed. In fact, we can find verbs that either match the dot type perfectly (i.e. select

it), or exploit one of its constituents, with the latter actually representing the large majority of cases.

| **selection** | |
|---|---|
| leggere (“read”) un libro / lettera / etc. | |
| **dot-exploitation** | |
| *phys* | bruciare (“burn”), portare (“carry”) un libro / imbucare (“post”), distruggere (“destroy”), raccogliere (“pick up”) una lettera / posare (“put down”), distribuire (“distribute”) un giornale / conservare (“keep”) un messaggio. |
| *info* | amare (“love”), citare (“quote”) un libro / riassumere (“summmarize”), comprendere (“understand”) una lettera / correggere (“correct”), conoscere (“know”) un articolo / censurare (“censor”), discutere (“discuss”) un testo / riempire (“fill”), commentare (“comment”) un giornale / ripensare (“rethink”) contestare (“dispute”) un discorso / commentare (“comment”) un’intervista. |
| *event* | riprendere (“start again with”), concludere (“conclude”), improvvisare (“improvize”), troncare (“cut”), un discorso / iniziare (“start”), interrompere (“stop”) un’intervista. |
| *organization* | danneggiare (“damage”), dirigere (“direct”), lasciare (“leave”), obbligare (“force”), il giornale |

Table 5 – Semantic operations in the verbal LSs

Interestingly, data also tell us that there are significant differences as to how frequently the single constituents of a dot-type are exploited: for instance, both *articolo* and *testo* combine much more frequently with i*nfo*-selecting verbs rather than with *phys*-selecting verbs, while they co-occur with *phys*-selecting verbs less frequently than *libro* and *lettera* (cf. Table 6).

| **articolo** | |
|---|---|
| *phys* | firmare (“sign”), spostare (“move”) |
| *info* | approvare (“approve”), bocciare (“reject”), citare (“quote”), votare (“vote”), correggere (“correct”), ignorare (“ignore”), commentare (“comment”), conoscere (“know”) |
| **testo** | |
| *phys* | firmare (“sign”), perdere (“lose”) |
| *info* | approvare (“approve”), votare (“vote”), conoscere (“know”), analizzare (“analyze”), presentare (“present”), revisionare (“amend”), discutere (“discuss”), censurare (“censor”), citare (“quote”), decifrare (“decipher”), difendere (“defend”), spiegare (“explain”), controllare (“check”) |
| **libro** | |
| *phys* | bruciare (“burn”), mandare (“send”), portare (“carry”) |
| *info* | amare (“love”), citare (“quote”), studiare (“study”) |
| **lettera** | |
| *phys* | imbucare (“post”), conservare (“keep”), infilare (“put”), distruggere (“destroy”), raccogliere (“pick up”), esibire (“exhibit”), ritrovare (“find again”), perdere (“lose”), portare (“bring”) |
| *info* | censurare (“censor”), scorrere (“scroll”), riassumere (“summmarize”), interpretare (“interpret”), esaminare (“examine”), comprendere (“understand”), spiegare (“explain”), ricordare (“remember”) |

Table 6 – Asymmetries in dot-exploitations

These asymmetries clearly bring afore the theoretical question whether these types should be considered dots or if they should rather be regarded as tensor types that are coerced contextually.

Finally, the analysis suggests that there are differences among the various nouns with respect to the encoding of the medium of the information. For instance, *testo* unlike *libro* combines easily both with verbs selecting for the written dimension (e.g. *leggere*) and with verbs selecting for the sound dimension (e.g. we find *ascoltare, cantare un testo* but not *ascoltare un libro*). We could then ask ourselves if it would not be more appropriate to consider *testo* as belonging to the type *info* and assume that the physical dimension is coerced contextually.

LSs also reveal some more complex examples, such as for instance *accusare un libro*. In fact, one does not really accuse a book, but rather the person who wrote it. Therefore, this case appears to be an instance of coercion via introduction of the type *human*. The same holds true for *difendere un testo* “defend a text”, *condannare una lettera* “condemn a letter”, etc. If so, it appears that dot-types like *book* do not only compose by selection or exploitation, but can also themselves be coerced into a different type by introduction. This may be a clue that the interplay between the type system and the compositional operations is more complex than the one depicted in Table 1. Additional examples of coerced dot-types are *leggere un discorso*, *pubblicare un’intervista*, *consegnare una dichiarazione*. In these cases, the physical dimension is introduced, which is not part of the inherent denotation of these nouns.

Next to domain-preserving operations as the ones discussed above, the data also bring up examples of coercions across domains (Pustejovsky 2001), like the ones reported below:[1]

[1] Remember that operations across domains are not included in Table 1.

| | |
|---|---|
| *libro* | ambientare (“set”) terminare (“finish”), cominciare (“start”) |
| *romanzo* | finire (“finish”), cominciare (“start”) |
| *articolo* | concludere (“conclude”), iniziare (“start”), cominciare (“begin”), terminare (“finish”), chiudere (“close”) |
| *testo* | completare (“complete”), finire (“finish”) |
| *lettera* | concludere (“conclude”), terminare (“finish”), interrompere (“interrupt”), finire (“finish”) |
| *messaggio* | concludere (“finish”), cominciare (“start”), finire (“finish”) |

Table 7 – Domain-shifting introduction of events

In order to account for coercions across domains (involving dot objects), we need to postulate an ordered sequence of compositional operations. First, an event is introduced through predicate selection: secondly, the Agentive and/or Telic specifications of the qualia structure of the nouns are exploited.

Coming now to prediction ii.), we can test it by analyzing the nominal LS of *leggere*, as a prototypical case of dot-selecting predicate. Again, the prediction is essentially confirmed by the data, with introduction working side by side to selection as the typical compositional operations of this predicate. An operation of exploitation is also detected (dot exploitation), occurring when the constituents of the dot-type of the noun match only partially the constituents of the dot-type selected by the predicate, as in *leggere il giornale*, where both the types *phys* and *info* are exploited, but not *organization*.

| | |
|---|---|
| **selection** | |
| leggere un libro (“book”), un articolo (“article”), un romanzo (“novel”), una lettera (“letter”) | |
| **dot-exploitation** | |
| leggere un giornale (“newspaper”) | |
| **introduction** | |
| *phys*: | leggere la trama (“plot”), la musica (“music”), un film (“movie”), un discorso (“speech”) |
| *info* | leggere la mano (“hand”), leggere una lapide (“headstone”), un dispositivo (“device”), un contatore (“meter”) |
| *phys and info* | leggere l’anima (“soul”), gli umori (“mood”) |

Table 8 – semantic operations in the nominal LS of *leggere*

As for introductions, in some cases (*leggere la trama, la musica*) the verb introduces a physical, written dimension, while in others (*leggere la mano, il contatore*) a physical artifact is coerced into an entity endowed with informational content. Finally, in a number of instances (*leggere l’anima*, *gli umori*), both the physical and the informational dimensions seem to be simultaneously wrapped around the argument by the predicate. Notice, however, that the interpretation of these last examples is complicated by the fact that, as we already clarified in §. 4, next to activating typing operations, *leggere* itself can undergo co-compositions with the argument and licence new senses. In these last examples, for instance, we could assume that the meaning of *leggere* differs from the one it exhibits in *leggere il libro* etc. (=come to know the info contained in a physical object), and is close to a more abstract sense of interpreting, decoding, etc. Thus, instead of the verb introducing a physical dimension onto the nouns, the latter would act on the reverse way, co-composing with the verb to determine its specific sense in context. The corpus provides other even clearer instances of co-composition, such as *leggere una radiografia* (= interpret) and *leggere una favola a un bambino* (= talk it loud).

These facts might suggest that the problem of disambiguating between coercions and co-compositions is a truly theoretical issue that can not be directly answered by looking at distributional evidence in a corpus only. Corpus analysis could provide us with quantitative data concerning the distribution in contexts of a specific sense of a predicate. On other hand, a clear understanding of the differences between co-compositions and coercions will require that other factors are taken into account as well, such as for instance the computational costs that are associated with different compositional operations (e.g. introductions being more costly then exploitations).

## 7 Final remarks and future research

Although we are aware that we have barely scratched the surface of the complex organization of even the small lexical fragment that we presented above, we think we can conclude that the combinatorial analysis of LSs is a promising method to integrate type system inquiry with corpus processing. So far, we can say that this technique has allowed us to: i) confirm our assumptions about what the semantic type of a given N is; b) refine the representation of the qualia structure of N; c) investigate empirically operations of coercion and co-composition. At a more general level, the results of our research confirms the possibility establishing a virtuous circle of mutual feeding between corpus analysis and GL. Infact, on the one hand, GL mechanisms

to generate structured types represent a highly expressive theoretical framework that is able to account for the different behaviour of lexical items as emerging from their distributions in syntagmatic contexts. On the other hand, data-driven analysis can profitably be used to anchor type distinctions and modifications to corpus evidence.

From the methodological point of view, a key point in our argument is that the reconstruction of how the meaning of a V-arg combination is compositionally generated can not dispense from a preliminary analysis of the composing lexical items as far as their types and type structure are concerned. In GL, coercion phenomena and STs definition are actually *two sides of the same coin*. Coercion acts on the enriched structure of the semantic types and consists of operations of selection or expansion of the ST. On the other hand, STs are defined in terms of the potentiality they offer to trigger coercion phenomena in compositional processes. Thus, it is crucial to build a model of what is stored in the lexicon and how it is stored in order to represent how this information enters into compositional processes. This obviously does not exclude that the analysis of syntagmatic contexts to identify compositional operations will in turn feedback on the representation of the types themselves. In fact, one can always go back and remodel the structure of the type system harmonizing it with the result of the investigation of its compositional behaviour.

In the future we plan to greatly refine the notion of syntagmatic context, extending it to cover other arguments as well (first of all subjects), adjectival modifiers of argument nouns, adverbs, etc. and to expand the analysis to other semantic types making use of the methodology described here.

## References


Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, A., Mazzoleni, M., (2004), "Introducing the *"la Repubblica"* corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian", *Proceedings of LREC 2004*, Lisboa.

Bartolini, R., Lenci, A., Montemagni, S. and Pirrelli V. (2004), "Hybrid Constraints for Robust Parsing: First Experiments and Evaluation", *Proceedings of LREC 2004*, Lisboa.

Bouillon, P., Claveau, V., Fabre, C. and Sebillot, P. (2002), "Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method", *Proceedings of LREC 2002*, Las Palmas.

Dunning, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics*, 19 (1): 61-74.

Fellbaum, C. (ed.), (1998), *WordNet: An Electronic Lexical Database*, Cambridge MA: MIT Press

Hanks, P. (2006), "The Organization of the Lexicon: Semantic Types and Lexical Sets", *Proceedings of XII Euralex*, Turin.

Hanks, P. and Pustejovsky, J. (2005), "A Pattern Dictionary for Natural Language Processing' in *Revue française de linguistique appliquée,* 10 (2).

Yamada, I. and Baldwin, T. (2004), "Automatic Discovery of Telic and Agentive Roles from Corpus Data", in Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18), Tokyo, Japan: 115–26.

Pustejovsky, J. (2001), "Type Construction and the Logic of Concepts", in P. Bouillon and F. Busa (eds.), *The Syntax of Word Meaning*, Cambridge University Press, Cambridge.

Pustejovsky , J. (2007), "Type theory and Lexical Decomposition", in P. Bouillon and C. Lee (eds) *Trends in Generative Lexicon Theory*, Kluwer Publisher (in press).

Pustejovsky, J., Hanks, P., and Rumshisky. A. (2004), "Automated Induction of Sense in Context", *Proceedings of COLING 2004*, Geneva, Switzerland.

Pustejovsky, J., Havasi, C., Littman, J., Rumshisky, A., and Verhagen, M. (2006), "Towards a Generative Lexical Resource: The Brandeis Semantic Ontology", *Proceedings of LREC 2006*, Genoa, Italy.