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

Elisabetta Jezek
Department of Theoretical and Applied Linguistics, University of Pavia, Italy
jezek@unipv.it

Alessandro Lenci
Institute for Computational Linguistics, Centro Nazionale Ricerche, Pisa, Italy
alessandro.lenci@ilc.cnr.it

1. Introduction

In this paper we present an analysis of corpus-derived V-arg combinations aiming to provide a data-driven characterization of Lexical Types (LTs) and represent how types behave compositionally, i.e. how they enter compositional processes and are modulated by them. We will do so using the enriched compositional rules and the type system as presented in Pustejovsky (2006). Our main concerns are twofold: i.) first of all, we want to show with a specific case-study (§. 5 onwards) how a data-driven investigation can shed light on the structure and the combinatorics of LTs; 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.

2. Theoretical framework

One of the major developments of the GL theory in recent years has been the integration of the three level type system (*simple-*, *unified-*, *dot-types*) 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, 2006). 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 allows for 9 possibilities as far as operations on types go, as reported in Table 1 in the Appendix. Next to operations on types (triggered by the verb), GL compositional processes also include *co-composition* phenomena triggered by the arguments, which may license new interpretations of the predicate in context. Since both operations of *typing* and of *co-composition* may operate simultaneously on the same syntagmatic context, it follows that 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.

3. Why and how corpus evidence is 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). Word senses for verbs are disambiguated through corpus-derived syntagmatic patterns mapped to GL as a linguistic model of interpretation, which guides and constrains the induction of senses from word 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 LTs.

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 σ a given predicative complex extracted form a corpus: this can be represented by whatever syntagmatic structure, although for the sake of this paper we will restrict σ to V-arg combinations. Therefore we will assume σ to be V-N pairs such as $<eat–cake_{obj}>$ or $<read–book_{obj}>$, etc. Each σ is therefore a piece of observed evidence of the behavior of lexical items in context. The epistemological issue is thus the following: what kind of inferences we can draw from the extracted contexts σ about the lexical system?

Given a certain context σ 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 of the argument of the predicative item V? iii.) what is the particular operation that allowed N and V to compose semantically in σ? Our working hypothesis is precisely that these three questions can be answered by investigating the full combinatorial distributions of V and N in a corpus. 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 LTs 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 LTs 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 σ has been generated by the combinations of two different factors: i.) the structure of the LTs to which V and N in σ 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 σ. If σ represents our empirical observational datum, i.) and ii.) are the two *hidden parameters* that we have to discover. As we said above in §. 2, 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 enriched type system and the array of semantic operations that we have to assume as being operative in every syntagmatic context.

4. Corpus processing and data extraction

In this research we focus our attention on Italian data, although we believe that most of these claims extend to other languages quite straightforwardly. Our dataset includes 877,352 syntagmatic contexts σ 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 σ has been automatically extracted from a 20 million subset of the *La Repubblica Corpus,* a 450 million word corpus of written Italian newspaper articles. 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) most naturally and typically occurs as direct object. 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 σ 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 systems, as our analysis below will show.

5. Anatomy of a type: the case of `leggere` "read"

First of all, why *leggere*? The reason of choosing this verb as the starting point for our case study 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 2006: 29).

Consequently, given the battery of semantic operations illustrated in §. 2 above, pure selection will 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 §. 4 can be used to carry out a sort of "autoptic analysis" of types in order to evaluate whether out intuition about the selecting environments of the internal argument of *leggere* can be validated and simultaneously refined with the help of text-driven data. To this purpose, we extracted from out 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 in the Appendix we have reported the top 40 nouns of this nominal LS, ordered by decreasing log-likelihood (ll) values. As we said above, from the fact that a noun occurs in the nominal lexical set of *leggere* we can not simply infer that the type of the noun is *phys* ● *info*. This is accounted for by the basic assumption of GL that selection is not the only way lexical items compose, and that a noun not fitting the predicate selecting environment can be coerced into it either by selecting a component of its original type or by introducing a new type that is "wrapped" around the noun's one. This is immediately evident if we consider the case 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. What we would like to claim is that actually this problem arises also with other members of the same nominal LS. In other terms, we are faced here with a truly general methodological issue, i.e. *what does the fact of observing a given noun within the lexical set of a verb can 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 only by exploring the verbal LSs of the nouns that belong to the nominal LS of *leggere*. This actually means that we have to inspect a larger area of the combinatorial space of lexical items: i.e. we can try 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_k combines, with N_k a member of the nominal LS of V. For the case of *leggere*, we have extracted the verbal LS of a subset the nouns in Table 2. For reasons of space, we have reported in Table 3 in the Appendix only the top 10 verbs (ordered for decreasing values of log-likelihood values) of the verbal LSs of 10 nouns. These verbal LSs bring afore interesting regularities:

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

- a second set of nouns - *lettera* "letter", *messaggio* "message" - is also characterized by verbal LSs dominated by verbs referring to the physical and 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 analysis, we can conclude that the nominal LS of *leggere* considered so far actually reveals itself as an articulated and variegated space with respect to the semantic properties of its members. However, GL theory of semantic types can provide the right interpretive key to for such distributional facts. In fact, we can express the semantic properties of the nouns in Table 3 with the following type representation (using the notation of tensor types in Pustejovsky 2006):

(1) a. *libro* "book", *articolo* "article", *romanzo* "novel"
: *phys* ● *info* $\bullet_{\text{Telic}}$ READING_EVENTS {*read*, *reread*,…} $\bullet_{\text{Agentive}}$ WRITING_EVENTS {*write*, *modify*, …} $\bullet_{\text{Agentive}}$ PUBLISHING_EVENTS {*publish*, *print*, …}

b. *lettera* "letter", *messaggio* "message"
: *phys* ● *info* $\bullet_{\text{Telic}}$ READING_EVENTS {*read*, *reread*,…} $\bullet_{\text{Telic}}$ TRANSMISSION_EVENTS {*send*, *circulate*, *deliver*…} $\bullet_{\text{Agentive}}$ WRITING_EVENTS {*write*, *modify*, …} $\bullet_{\text{Agentive}}$ PUBLISHING_EVENTS {*publish*, …}

The differences among these lexemes can thus be captured in terms of their qualia structure. This also closely corresponds to most natural intuition 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 LS a number of verbs referring to 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.

6. Discovering lexical types

Besides allowing a refined representation of the nouns as far as their qualia structure is concerned (§. 5), the investigation of the verbal LSs also allows us to verify empirically our assumptions about what the type itself associated to these nouns is, - both in terms of its level (say if it is a *simple type*, a *tensor type* or *dot type*) and its constitution (say if it is *phys* ● *info* or *event* ● *info*). However, since verbs perform different kinds of operations depending on the type they meet, when we explore a verbal LS with the aim of reconstructing the type of the noun and distinguishing it from other possible types, we need to distinguish between the *best verbs* for N and the *coercing verbs*. Best verbs are those verbs that match the noun type up to the point of entering its qualia structures. Remember, in fact, that agentive and telic qualia are in fact the protopical events related to the noun's fucntionality or mode of creation. Coercing verbs are instead verbs that exploit the original type of a noun and cause it to become the expected one. As we already pointed out in §. 5, the problem is that within the most frequent σ, we can expect to find cases of pure selection (perfect matching between selecting and selected type: so, best verbs), exploitation (failure of matching:

coercing verbs) or introductions (coercing verbs; although we assume introductions to be more likely situated in low frequencies of σ).

Consider again the nouns *libro* "book", *articolo* "article", *testo* "testo", *romanzo* "novel", *lettera* "letter", *messaggio* "message", and compare them to *giornale* "newspaper", *intervista* "interview", *dichiarazione* "declaration", *discorso* "discourse". As we can see from Table 4 in the Appendix, 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, suggesting that they do not belong to the same type. Remember that by inspecting the verbal LSs of *libro, romanzo, lettera* etc. we observed above that these nouns typically co-occur with *phys*●*info*-selecting verbs (*leggere*, *rileggere*) or alternatively with *phys*-selecting verbs (*sfogliare, risfogliare, portare*) or *info*-selecting verbs (*pubblicare, presentare)*. If we turn to 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" (next to *phys*-selecting verbs like *posare* and *info*_selecting verbs like *criticare*, *censurare*, that also appear in the full LS of *giornale*) - 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 members lexemes discussed in §. 5. 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*).

Let us now look at the verbal LS of *dichiarazione* "declaration", *discorso* " speech" and *intervista* "interview" in Table 4. What immediately comes into sight is that the physical and/or printed dimension is totally in the background, if not lacking: although these nouns co-occur with *phys*-selecting verbs, they more often combine with verbs that refer to the oral/sound dimension (e.g. *pronunciare*, *ascoltare*, *rilasciare*, *registrare*, etc.) or to the eventive, time enduring character of the entities to which the nouns refer to (e.g. to event-selecting verbs like *concludere*, *riprendere*, ecc.). 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 speeches and declarations can not be written, but rather that this might not be the most salient dimension for these nouns. 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 §. 5, we can express the semantic properties of these nouns with the following type representation (using the notation of tensor types in Pustejovsky 2006):

(2) *dichiarazione* "declaration", discorso "speech", intervista "interview"
: *event* ● *info* $\bullet_{\text{Agentive}}$ SPEECH_EVENTS {*pronounce*, *address*, *give a speech*…} $\bullet_{\text{Telic}}$ LISTENING_EVENTS {*listen*, …}

To sum up, from the analysis of the verbal LSs carried out in §. 5 and 6, we may conclude that the variations in the verbal LSs can be interpreted as an indicator of two main facts: what is the specification in the QS of a noun, and what is the level and the nature of its 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.

7. An overall map of compositional operations

Besides allowing us to confirm or contradict our assumptions about lexical types and their structure, corpus analysis can help us to improve our understanding of how types behave compositionally, i.e. how they enter compositional processes and are modulated by them, and thus to contribute to reconstruct how the meaning of a VN combination is generated. 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 esponentially the creative ways in which we can use them to express meanings. This assumption raises a lot of questions: if so, what are the modalities in which coercion takes place? What are the rules and restrictions? Can all lexical items acrivate coercions? What aspect(s) of the type are more likely to be coerced?. Remember that Pustejovsky (2006) assumes that i. it is the predicates that select the type of their arguments according to their inherent semantic properties and impose various kinds of coercions on that type if it does not correspond to their selectional restrictions; ii. arguments may simoultaneously modify the semantics of the predicate by co-composition. We would like to claim that it is precicely these assumtions 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 reference of our analysis, we see that the GL organization of the type system makes two specific predictions concerning the compositional modes of dot-types: 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 some of the nouns above, that as a result of our analysis in §. 5 and 6 have been assigned either to the *phys* • *info* type (e.g. *libro*, *romanzo*, *lettera*, etc.) or to the *event* • *info* type (e.g. *discorso*, etc.), 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 type perfectly (i.e. select it), or exploit parts of its dot constitution, with the latter actually representing the large majority of cases:

(3)

selection:	leggere (read) un libro / lettera / etc.
exploitation:	a. phys – bruciare (burn), portare (carry) il libro; imbucare (post), infilare (put), distruggere (destroy), raccogliere (pick up) la lettera; posare (put down), distribuire (distribute) il giornale; conservare (keep) il messaggio. b. info – citare (quote) un libro, riassumere (summmarize), comprendere (understand) la lettera; correggere (correct), conoscere (know) l'articolo; censurare, discutere (discuss) un testo; riempire (fill in), commentare (comment) il giornale; ripensare, contestare il discorso; commentare l'intervista. c. event – riprendere (start again with), concludere (conclude), improvvisare (improvise), troncare (cut) il discorso; iniziare (start), interrompere (stop), vedere (see) un'intervista

d. organization –danneggiare (damage), dirigere (direct), lasciare (leave) il giornale.

Interestingly, data also tell us that the constituents of a dot type are often not selected to the same degrees: for instance, both *articolo* and *testo* combine more frequently with i*nfo*-selecting verbs rather than with *phys*-selecting verbs. The same holds true for *articolo*, *testo* and *romanzo* that present a lower number of *phys*-selecting verbs than *libro* and *lettera*. This suggests that some dot types might be asymmetric as far as their internal constitution goes, or, alternatively, be in fact tensor types that are coerced contextually. Finally, not all nouns seems to specify the way 'information' is coded. For instance, *testo* unlike libro can easily combine both with verbs referring to the written dimension (e.g. *leggere*) as well as with verbs that refer to 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 simply *info* and assume that the physical dimension is coerced contextually.

LSs also reveal some more complex examples, such as for instance *accusare un libro*, *ambientare un libro*, etc. In fact, you really do not accuse a book, but rather the person who wrote it, and *ambientare* refers to the setting of the events and situations described in the book. Therefore, both cases appear to be clear instances of coercion via introduction. The same holds true for *difendere un testo* "defend a text" and *condannare una lettera* "condemn a letter" etc. If so, 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 much more complex than the one depicted in table 1.

Coming 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:

(4)

selection:	leggere un libro / lettera / lapide etc.
introduction:	a. phys – leggere la trama, la musica, un film, il discorso
	b. info – leggere la mano, leggere il dispositivo, il contatore
	c. phys and info – leggere l'anima, leggere gli umori, leggere i segni

In some cases the verb *leggere* introduces a physical, written dimension, while in others a non-semiotic artifact is coerced into an entity endowed with informational content. Finally, in a number of instances, both dimensions seem to be simultaneously wrapped around the argument by the predicate. Actually, the latter appears to be a boundary case between introduction and co-composition, since *leggere* acquires a different interpretation than the standard one, close to more abstract senses of interpreting, decoding, etc. Thus, instead of the verb introducing a physical dimension onto the nouns, the latter act on the reverse way, co-composing with the verb to determine its specific sense in context. A similar situation holds for *vedere la lettera* "see the letter": the verb selects the type *phys*, but since *lettera* is *phys* • *info* as a consequence of co-composition *vedere* is reinterpreted as 'understand'.

8. 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. 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 modification to corpus evidence. From the methodological point of view, it is important to remark that the reconstruction of how the meaning of a V-arg combinations 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 fact, in GL coercion phenomena and LTs definition are 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 LT. On the other hand, LTs 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 try to explain 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, for instance extending it to cover other arguments as well (first of all subjects), adjectival modifiers of argument nouns, adverbs, etc. and to extent the analysis to other types making use of the mothodology described here.

References

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. (2006), "Type Theory and Lexical Decomposition" in ??

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.

Appendix

	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 in GL

noun	ll value
libro “book”	225,44
giornale “newspaper”	174,98
articolo “article”	133,28
lettera “letter”	96,77
romanzo “novel”	76,63
testo “text”	58,34
documento “document”	56,42
intervista “interview”	52,37
comunicato “communiqué”	49,23
dichiarazione “statement”	48,07
pagina “page”	47,76
sceneggiatura “script”	44,17
riga “line”	42,03
discorso “speech”	41,07

noun	ll value
cartella “page”	40,64
messaggio “message”	36,10
relazione “report”	35,14
passo “passage”	34,60
resoconto “report”	30,04
parola “word”	29,71
frase “sentence”	28,75
sentenza “sentence”	25,93
motivazione “justification”	23,39
Freud	19,96
Financial Times	19,40
omelia “sermon”	16,92
notizia “news”	16,14
saggio “essay”	16,04

noun	ll value
missiva “missive”	15,85
telegramma “telegram”	14,97
poesia “poem”	14,77
verdetto “verdict”	14,62
brano “passage”	14,62
nota “note”	14,51
opera “work”	14,20
Rimbaud	14,19
sofisma “sophisma”	14,19
Tuttosport	14,19
scritta “writing, notice”	11,75
telex “telex”	11,59

Table 2 – top 40 nouns in the LS of `leggere`

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

Table 3 – top 10 verbs in the LS of a set of nouns of the LS of `leggere`

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`