Qualia Structure and the Compositional Interpretation of Compounds

Michael Johnston[§] and Federica Busa[†]

Research Lab for Linguistics and Computation,
Computer Science Department,
Volen Center for Complex Systems,
Brandeis University,
Waltham, MA 02254

[§]johnston@cs.brandeis.edu [†]federica@cs.brandeis.edu

Abstract

The analysis of nominal compound constructions has proven to be a recalcitrant problem for linguistic semantics and poses serious challenges for natural language processing systems. We argue for a compositional treatment of compound constructions which limits the need for listing of compounds in the lexicon. We argue that the development of a practical model of compound interpretation crucially depends on issues of lexicon design. The Generative Lexicon (Pustejovsky 1995) provides us with a model of the lexicon which couples sufficiently expressive lexical semantic representations with mechanisms which capture the relationship between those representations and their syntactic expression. In our approach, the qualia structures of the nouns in a compound provide relational structure enabling compositional interpretation of the modification of the head noun by the modifying noun. This brings compound interpretation under the same rubric as other forms of composition in natural language, including argument selection, adjectival modification, and type coercion (Pustejovsky (1991,1995), Bouillon 1995). We examine data from both English and Italian and develop analyses for both languages which use phrase structure schemata to account for the connections between lexical semantic representation and syntactic expression. In addition to applications in natural language understanding, machine translation, and generation, the model of compound interpretation developed here can be applied to multi-lingual information extraction tasks.

1 Introduction

The analysis of nominal compound constructions has proven to be a recalcitrant problem for linguistic semantics (Bergsten 1911, Jespersen 1942, Marchand 1970, Lees 1970, Downing 1977, Levi 1978, Warren 1987), and their analysis has presented a serious challenge for natural language processing systems (Finin 1980, McDonald 1982, Isabelle 1984, Alshawi 1987, Hobbs et al 1993, Bouillon et al 1992, Jones 1995, Johnston, Boguraev, and Pustejovsky 1995). In this paper, we argue that composition in compound constructions involves specification of the arguments of predicate structures within the qualia structure of the head noun. In essence, the qualia structure provides the 'glue' which links together the semantic contributions of modifying nouns and the head noun in the compound. This approach has significant advantages over approaches which introduce abstract relations to describe the form of modification involved (Levi 1978, Jones 1995), in that it assimilates the composition of complex nominals to other compositional processes of natural language. The predicates in the qualia are not there just to account for compounds but also to account for a wide variety of forms of composition and interpretation including argument selection, adjectival modification, and type coercion (Pustejovsky (1991,1995), Bouillon 1995). In order to account for the availability of compounds and other complex nominals we utilize phrase structure schemata which capture the different ways in which head nouns and modifying nouns can compose. We develop and illustrate our approach using data from English and Italian. Consider the correspondences below in (1).

(1)	a.	bread knife	b.	wine glass	c.	bullet hole
		coltello da pane		bicchiere da vino		foro di pallottola
	d.	lemon juice	e.	glass door	f.	silicon breast
		succo di limone		porta a vetri		seni al silicone

English compounds consist of sequences of nouns. The final noun is the head and the others are modifiers. Italian, on the other hand, generally avoids compounds. Most of the compounds that are possible involve an underived verb (e.g. *portadocumenti* (document holder), *segnalibro* (bookmark), *spaventapasseri* (scarecrow), etc.)[1] In place of the compound construction, Italian utilizes post-modifying prepositional phrases to construct complex nominals.

Consider first the forms in (1a) and (1b). In both of these forms the modifying noun provides information regarding the purpose or function of object described by the head noun. In (1a), *bread* specifies the object which the knife is typically used to cut. In (1b), *wine* specifies the substance that the glass is used to hold. For forms in which the modifier specifies some aspect of the purpose of the head, the preposition in the Italian form generally is *da.*

In (1c) and (1d), the modifier relates to the origin of the object described by the head noun, how it was brought about. A *bullet hole* is a hole which was brought about by the passage of a bullet, and *lemon juice* is juice that is brought about by squeezing a lemon. For forms in which the modifier specifies some aspect of the origin of the head, the appropriate preposition for the Italian form appears to be *di.*

In (1e) and (1f), the modifier relates to the constitution of the object described by the head noun, what it is made of. A *glass door* is a door made of glass, while a *silicon breast* is a breast, at least partially, composed of silicon. For forms in which the modifier specifies the constitution of the head, the appropriate preposition in Italian is *a.*

In English, the form of the semantic relation that holds between the modifying noun and the

[1] See Beard(1996) for an interesting account of why Italian does not utilize the compound construction.

head noun is unspecified and left implicit, while in the corresponding Italian complex nominals, it is partially specified by the preposition which introduces the modifying noun. This difference between English and Italian is representative of a more general difference in structure between Germanic languages and Romance languages. Complex nominal expressions in Germanic languages frequently involve compounding while complex nominals in Romance languages use post-modifying prepositional phrases. One motivation for examination of parallel data from English and Italian is that it can reveal the otherwise implicit relation between the elements of an English compound. This difference in the internal structure of nominals has important consequences for multilingual language processing. For example, in order to translate English complex nominals into Italian you need to determine the nature of the semantic relation in the compound in order to determine the preposition to use in the Italian form. We return to applications of this work later in Section 7. Our analysis of compounds in English and complex nominals in Italian utilizes the representational framework of the Generative Lexicon (GL) (Pustejovsky 1991,1995). The next section provides a brief sketch of this approach.

2 The Generative Lexicon

For the purposes of this paper, we will simplify the representational structure of a GL lexical entry to include four levels of representation: *type structure*, *argument structure*, *event structure* (for verbs), and *qualia structure*. The latter in turn expresses four aspects of the meaning of the lexical item: FORMAL, CONSTITUTIVE, TELIC, and AGENTIVE. These lexical entries are encoded using typed feature structures. The basic layout of the lexical entries we employ is given in (2).

$$(2)\quad \begin{bmatrix} \alpha & \\ \text{TYPESTR} = & \begin{bmatrix} \text{ARG1} = \textbf{the type of } \alpha \end{bmatrix} \\ \text{ARGSTR} = & \begin{bmatrix} \text{D-ARG1} = \textbf{other arguments in the qualia} \end{bmatrix} \\ \text{EVENTSTR} = & \begin{bmatrix} \text{E1} = \textbf{events in the qualia} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{FORMAL} = \textbf{isa-relation} \\ \text{CONSTITUTIVE} = \textbf{parts of } \alpha \\ \text{TELIC} = \textbf{purpose of } \alpha \\ \text{AGENTIVE} = \textbf{how } \alpha \textbf{ is brought about} \end{bmatrix} \end{bmatrix}$$

Given this model of lexical representation a noun such as *knife* has the entry in (3). The predicates in the qualia specify the definitional properties of *knife*. Participants in these predicates other than the knife itself are listed as default arguments (D-ARG1, D-ARG2, and D-ARG3) in

ARGSTR.

(3)
$$\left[\begin{array}{ll} \textbf{knife} & \\ \text{TYPESTR} = & \left[\text{ARG1} = \boxed{x}\,\textbf{artifact\_tool}\right] \\ \text{ARGSTR} = & \left[\begin{array}{l} \text{D-ARG1} = \boxed{y}\,\textbf{physobj} \\ \text{D-ARG2} = \boxed{w}\,\textbf{human} \\ \text{D-ARG3} = \boxed{z}\,\textbf{human} \\ \text{D-E1} = \boxed{e_1}\,\textbf{transition} \\ \text{D-E2} = \boxed{e_2}\,\textbf{process} \end{array}\right] \\ \text{QUALIA} = & \left[\begin{array}{l} \text{FORMAL} = \boxed{x} \\ \text{CONSTITUTIVE} = \{\textbf{blade,handle,...}\} \\ \text{TELIC} = \textbf{cut\_act}(\boxed{e_2},\boxed{w},\boxed{x},\boxed{y}) \\ \text{AGENTIVE} = \textbf{make\_act}(\boxed{e_1},\boxed{z},\boxed{x}) \end{array}\right] \end{array}\right]$$

Representations such as that in (3) are intended to be the values of a CONTENT attribute which specifies the semantic content of a lexical item. All of the representations for single words and complex nominals throughout the rest of the paper consist only of the value of CONTENT. The phrase structure schemata presented contain full lexical entries which have a CONTENT attribute as well as an ORTH and a DTRS attribute.

In the following three sections, we will show how the free classes of compounds considered so far can be treated as instances of telic, agentive, and constitutive qualia modification respectively.

3 Telic Qualia Modification

In order to illustrate our approach, we will start with examples such as *bread knife* (1a), in which the modifying noun relates to the purpose of the head noun. The preferred interpretation of this compound is that it is a knife which is used to cut bread. The fact that a knife is an object whose inherent purpose is to cut things is encoded by the predicate **cut_act** in the TELIC role (see (3) above). The function of the modifier *bread* is to specify the third argument of the **cut_act** relation. The feature structure associated with *bread knife* will be as in (4). The first default argument D-ARG1 has been specialized from **physobj** to **bread** and this value is structure-shared with the third argument position in the **cut_act** predicate.

(4)
$$\left[\begin{array}{ll} \textbf{bread knife} & \\ \text{TYPESTR} = & \left[\text{ARG1} = \boxed{x}\,\textbf{artifact\_tool}\right] \\ \text{ARGSTR} = & \left[\begin{array}{l} \text{D-ARG1} = \boxed{y}\,\textbf{bread} \\ \text{D-ARG2} = \boxed{w}\,\textbf{human} \\ \text{D-ARG3} = \boxed{z}\,\textbf{human} \\ \text{D-E1} = \boxed{e_1}\,\textbf{transition} \\ \text{D-E2} = \boxed{e_2}\,\textbf{process} \end{array}\right] \\ \text{QUALIA} = & \left[\begin{array}{l} \text{FORMAL} = \boxed{x} \\ \text{CONSTITUTIVE} = \{\textbf{blade,handle,...}\} \\ \text{TELIC} = \textbf{cut\_act}(\boxed{e_2},\boxed{w},\boxed{x},\boxed{y}) \\ \text{AGENTIVE} = \textbf{make\_act}(\boxed{e_1},\boxed{z},\boxed{x}) \end{array}\right] \end{array}\right]$$

In the GL representation, all of the participants which show up in the predicates in qualia are listed as default argument parameters in the ARGSTR.

In order to account for the availability of compound forms in English, we utilize a family of phrase structure schemata. These schemata are essentially the same kind of entity as the Immediate Dominance Schemata employed in Head-driven Phrase Structure Grammar (Pollard and Sag 1994). They are schemata which license the availability of complex nominals, which we treat as phrasal signs. These schemata are essentially phrase structure rules. Compounds are licensed and interpreted as part of the process of parsing.

The combination of words into compound forms could also be captured using lexical rules (Flickinger 1987, Pollard and Sag 1987). We have chosen to use phrase structure schemata rather than lexical rules on the basis of storage considerations. Each lexical rule used for compounds will license a great many modifiers for large number of potential heads. If the lexical rules are used at a pre-compilation stage in order to flesh out the lexicon, allowing lexical rules for compounds will result in a massive increase in the size of the lexicon. For each noun, a huge number of compound forms will be generated. If you allow lexical rules for compounds to apply at runtime during the parsing process, then the storage problem is avoided, but then they are really not any different from phrase structure schemata.

We will show the schemata as rules here. They can also be encoded as single feature structures. The basic structure of the schemata licensing the combination of nouns to form noun compounds is as in (5).

$$\begin{array}{llll} & \text{MODIFIER NOUN} & \text{HEAD} & \\ & \boxed{1}\begin{bmatrix}\text{ORTH} = \beta\end{bmatrix} & \boxed{2}\begin{bmatrix}\text{ORTH} = \alpha\end{bmatrix} & \Rightarrow \\ & \text{COMPOUND} & & \\ (5) & \begin{bmatrix}\text{ORTH} = \beta\alpha \\ \text{DTRS} = \begin{bmatrix}\text{HEAD} = \boxed{2} \\ \text{MOD} = \boxed{1}\end{bmatrix}\end{bmatrix} & & \end{array}$$

The schemata differ with respect to the constraints placed on the CONTENT values and the way in which the CONTENT values of the head and the modifier are composed to generate the CONTENT for the compound as a whole. The availability of compound forms such as *bread knife*, where the modifier specifies an argument in the TELIC, is accounted for by the schema in (6).

$$\begin{array}{llll} & \text{MODIFIER NOUN} & \text{HEAD} & \\ & \boxed{1}\begin{bmatrix}\text{ORTH} = \beta \\ \text{CONTENT} = \boxed{4}\textbf{individual}\end{bmatrix} & \boxed{2}\begin{bmatrix}\text{ORTH} = \alpha \\ \text{CONTENT} = \boxed{3}\end{bmatrix} & \Rightarrow \\ & \text{COMPOUND} & & \\ (6) & \begin{bmatrix}\text{ORTH} = \beta\alpha \\ \text{CONTENT} = \begin{bmatrix}\boxed{3} = \begin{bmatrix}\text{ARGSTR} = \begin{bmatrix}\text{D-ARG1} = \boxed{4}\end{bmatrix}\end{bmatrix}\end{bmatrix} \\ \text{DTRS} = \begin{bmatrix}\text{HEAD} = \boxed{2} \\ \text{MOD} = \boxed{1}\end{bmatrix}\end{bmatrix} & & \end{array}$$

In this notation, the structures describing semantic types are the values of an attribute CONTENT, and ORTH specifies the orthographic form. The CONTENT of the resulting compound is inherited from the head noun. In order to access the argument in the TELIC, the CONTENT value of the modifier is structure-shared with the first default argument in the CONTENT of the head. The modifying noun must be of semantic type **individual** and its CONTENT value is structure-shared with the D-ARG1 in the ARGSTR of the resulting compound. The lexical representation of the compound also contains an attribute DTRS containing a HEAD and a MOD value. These are structure-shared with the lexical representations for the head noun and the modifying noun respectively.

This schema is one of a number which are used to license this kind of modification of default arguments. There will also be schemata for modification of other default arguments. The fact that the CONTENT of the compound always comes from the head noun is captured by having all of the compound phrase structure schemata, which are themselves implemented as types, all inherit the constraint specified by the structure-sharing index $\boxed{3}$.

As we saw before, if the modifier specifies an argument in the TELIC qualia role, the preposition in Italian is *da*. In order to account for the Italian forms, as in the English case, we utilize phrase structure schemata. In this case, the schema (7) specifies that the sequence HEAD NOUN, *da*, MODIFYING NOUN can be interpreted as having the semantic content of the modifying noun specify one of the arguments within the TELIC role.

$$\overset{\text{HEAD}}{\boxed{2}\begin{bmatrix}\text{ORTH} = \alpha \\ \text{CONTENT} = \boxed{3}\end{bmatrix}}\ \text{da}\ \overset{\text{MODIFIER NOUN}}{\boxed{1}\begin{bmatrix}\text{ORTH} = \beta \\ \text{CONTENT} = \boxed{4}\textbf{individual}\end{bmatrix}} \Rightarrow$$

$$(7)\quad \begin{bmatrix}\text{COMPOUND} & \\ \text{ORTH} = \alpha\ da\ \beta & \\ \text{CONTENT} = \begin{bmatrix}\boxed{3} = \begin{bmatrix}\text{QUALIA} = \begin{bmatrix}\text{TELIC} = \mathbf{R}[..\boxed{4}..]\end{bmatrix}\end{bmatrix}\end{bmatrix} & \\ \text{DTRS} = \begin{bmatrix}\text{HEAD} = \boxed{2} \\ \text{MOD} = \boxed{1}\end{bmatrix} & \end{bmatrix}$$

The indeterminacy with respect to which argument in the TELIC is coindexed with the modifier in schema (7) is a shorthand representation. A number of phrase structure schemata are used, each specifying linking to a different argument position in the TELIC.

For Italian, the nature of the modification can alternatively be directly encoded in the lexical entry for the preposition. The composition could then licensed by a more general phrase structure schema which would work with all of the different prepositions.

4 Agentive Qualia Modification

Compounds such as *bullet hole* and *lemon juice* (1 c,d), in which the modifier relates to the origin or bringing about of the object described by the head noun, are treated as modification of the AGENTIVE role. In the case of *lemon juice*, the head *juice* will have a **squeeze_act** as its AGENTIVE and the object squeezed will be listed as a default argument. The function of the modifying noun *lemon* is to further subtype this argument. This is possible because **lemon** is a subtype of **fruit**. These English forms will be accounted for by another schema licensing default argument type specification, like that in (6) above. The resulting representation for *lemon juice* is as in (8). The corresponding forms in Italian utilize the preposition *di*. The Italian forms are accounted for by a schema like (8), except that the preposition is *di* and the linkage is to the AGENTIVE qualia role.

$$(8)\quad \begin{bmatrix}\textbf{lemon juice} & \\ \text{TYPESTR} = \begin{bmatrix}\text{ARG1} = \boxed{x}\textbf{liquid}\end{bmatrix} & \\ \text{ARGSTR} = \begin{bmatrix}\text{D-ARG1} = \boxed{y}\textbf{lemon} \\ \text{D-E1} = \boxed{e_1}\textbf{transition}\end{bmatrix} & \\ \text{QUALIA} = \begin{bmatrix}\text{FORMAL} = \boxed{x} \\ \text{AGENTIVE} = \textbf{squeeze\_act}(\boxed{e_1},\boxed{y},\boxed{x})\ldots\end{bmatrix} & \end{bmatrix}$$

5 Constitutive Qualia Modification

Another common function of modifiers in complex nominals is to specify a subpart of the denotation of the head noun or the material of which it is composed. Examples of this are given in (1 e,f). In our treatment, this involves modification of the CONSTITUTIVE role. The prepositions used in Italian for this sort of modification are *a* and *al*. The modifiers *glass* and *silicon* denote materials. When composed with nominals such as *door* and *breast* they specify elements of the CONSTITUTIVE role. For example, *glass door* is represented as in (9). These forms are licensed using further phrase structure schemata for English and Italian.

(9)
$$\left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{glass door}} \\
\text{TYPESTR} = & \left[\begin{array}{l} \text{ARG1} = \boxed{x}\,\textbf{phys\_obj} \\ \text{ARG2} = \boxed{y}\,\textbf{aperture} \end{array}\right] \\
\text{ARGSTR} = & \left[\begin{array}{l} \text{D-ARG1} = \boxed{w}\,\textbf{individual} \\ \text{D-ARG2} = \boxed{z}\,\textbf{individual} \\ \text{D-E1} = \boxed{e_1}\,\textbf{transition} \\ \text{D-E2} = \boxed{e_2}\,\textbf{transition} \end{array}\right] \\
\text{QUALIA} = & \left[\begin{array}{l} \text{FORMAL} = \textbf{hold}(\boxed{y},\boxed{x}) \\ \text{CONSTITUTIVE} = \{\boxed{x}\,\textbf{glass}\} \\ \text{TELIC} = \textbf{walk\_through\_act}(\boxed{e_2},\boxed{w},\boxed{y}) \\ \text{AGENTIVE} = \textbf{make\_act}(\boxed{e_1},\boxed{z},\boxed{x}.\boxed{y}) \end{array}\right]
\end{array}\right]$$

The basic pattern established so far is that modification of TELIC, AGENTIVE, and CONSTITUTIVE involves *da*, *di*, and *a*, respectively. This is a useful generalization but the correspondence between the different qualia roles and different choices of preposition in Italian is not as clear cut as this suggests. In the examples of TELIC qualia modification considered so far (1 a,b), the modifying noun was always of type **individual**. Matters become more complex when compounds in which the modifying noun describes an event are considered. These are addressed in the next section.

6 Telic Event Modifiers

In some forms where the modifier describes an event, the appropriate preposition in Italian is *da*, as in the forms in (10), while others the preposition is *di*, as in the forms in (11).

(10) a. hunting rifle
fucile da caccia
b. race car
macchina da corsa
c. carving wood
legno da intaglio

(11) a. destruction weapons
armi di distruzione
b. credit card
carta di credito
c. rest home
casa di riposo
d. concentration camp
campo di concentramento
e. divorce procedure
procedura di divorzio

In general, the TELIC use of the preposition *di* appears to select consistently for modifiers which denote events. Even though this does not yet explain the difference between (10) and (11), it already provides us with a restriction on the use of prepositions. In other words *da* selects for any type, while *di* is restricted to events. We assume the Vendlerian distinction between *activities*, *states*, *accomplishments*, and *achievements*. In addition, we adopt a decompositional view of event structure, as outlined in Pustejovsky (1991), in which the event structure representation of a lexical item makes reference to the configurational properties of subevents and arguments. In this framework, which allows us to make fine grained distinctions between event types, we can determine the selectional properties of *di* and *da*, on the basis of the event type of the modifiers. Nominals such as *hunting*, *race*, and *carving* describe activities. Nominals such as *destruction*, *credit*, and so on, in (11) above, describe the result of an activity. This distinction arises quite clearly in the glosses of (10) and (11). Compound forms such as *hunting rifle* or *race car* in (10), describe respectively an instrument which is used when hunting, and a vehicle that is driven for the purpose of racing. Conversely, the reading of the compounds in (11) makes explicit the result which is achieved by using a particular object. In particular (11a) refers to weapons that bring about destruction; (11b) to a card that brings about a credit, and so on.

Unlike the operation which derives *bread knife* by associating the modifier to an argument position in the TELIC role of *bread*, the compositional operations which involve events produce a more complex structure. We argue that compounds where the modifying noun describes an event, such as those in (10), involve co-composition of the qualia structures of the head and the modifier. The resulting representation has a complex TELIC role with “sub-qualia”. In the case of *hunting* rifle, the TELIC of *rifle*, which is **fire** provides the AGENTIVE within the TELIC of the compound. The modifier *hunting* is a process nominal and provides **hunt** as the TELIC within the TELIC of the compound. Through the application of phrase structure schemata which constrain this co-composition, we obtain the representation in (12) for *hunting rifle.*

(12)
$$\left[\begin{array}{l}
\textbf{hunting rifle} \\
\text{TYPESTR} = \left[\begin{array}{l}\text{ARG1} = \boxed{x}\,\textbf{rifle}\end{array}\right] \\
\text{ARGSTR} = \left[\begin{array}{l}\text{D-ARG1} = \boxed{w}\,\textbf{human} \\ \text{D-ARG2} = \boxed{z}\,\textbf{prey}\end{array}\right] \\
\text{EVENTSTR} = \left[\begin{array}{l}\text{D-E1} = \boxed{e_1^P}\,\textbf{process} \\ \text{D-E2} = \boxed{e_2^P}\,\textbf{process}\end{array}\right] \\
\text{QUALIA} = \left[\begin{array}{l}\text{FORMAL} = \boxed{x} \\ \text{TELIC} = \left[\begin{array}{l}\textbf{activity\_lcp} \\ \text{TELIC} = \textbf{hunt}(\boxed{e_2^P}, \boxed{w}, \boxed{z}) \\ \text{AGENTIVE} = \textbf{fire}(\boxed{e_1^P}, \boxed{w}, \boxed{x})\end{array}\right]\end{array}\right]
\end{array}\right]$$

The interpretation of the compound form *hunting rifle* can be glossed as follows:“a rifle which is used in its typical capacity (i.e. firing) for the purpose of performing the activity of hunting.” The assignment of a complex structure to an individual *quale* is coherent with the general interpretation of qualia structure. Exploiting these recursive properties of event-denoting qualia is not an ad-hoc move to account for the interpretation of complex nominals but is also motivated by the behavior of agentive nominals and their semantic contribution in context (cf. Busa 1996).

The modifying noun in Italian complex nominals with the preposition *di* describes the result that is achieved by performing the particular function associated with the head noun. The nominal *destruction*, in (11a), unlike the event nouns *hunting* and *race* which denote activities, is the nominalization of the transitional event denoted by the verb *destroy.* The two subevents, namely the process and the resulting state, in the event structure representation of the verb, are encoded in the nominalized form as separate events in the AGENTIVE and FORMAL roles, and they are related by the relation of temporal precedence $<_{\alpha}$. As argued in Pustejovsky (1995) this representation gives rise to the polysemous behavior of the nominal. It alternates between a process and a result interpretation. In *destruction weapon*, the embedded AGENTIVE in the TELIC is again the TELIC of the head *weapon*, and the embedded TELIC is the resulting state from the semantics of *destruction.*

The resulting TELIC is a **process-result-lcp**, as shown in (13).

$$(13)\quad \left[\begin{array}{ll} \multicolumn{2}{l}{\textbf{destruction weapon}} \\ \text{TYPESTR} = & \left[\begin{array}{l}\text{ARG1} = \boxed{\text{x}}\,\textbf{weapon}\end{array}\right] \\ \text{ARGSTR} = & \left[\begin{array}{l} \text{D-ARG1} = \boxed{\text{y}}\,\textbf{individual} \\ \text{D-ARG2} = \boxed{\text{w}}\,\textbf{human} \\ \text{D-E1} = \boxed{e^{S}}\,\textbf{:state} \\ \text{D-E3} = \boxed{e_1^{P}}\,\textbf{:process} \\ \text{REST} = \boxed{e_1^{P}} <_{\propto} \boxed{e^{S}} \end{array}\right] \\ \text{QUALIA} = & \left[\begin{array}{ll} \text{FORMAL} = \mathbf{x} & \\ \text{TELIC} = & \left[\begin{array}{l} \textbf{process-result-lcp} \\ \text{FORM} = \textbf{destruction\_result}(\boxed{e^{S}},\boxed{\text{y}}) \\ \text{AGENTIVE} = \textbf{fire}(\boxed{e_1^{P}},\boxed{\text{w}},\ \boxed{\text{x}}) \end{array}\right] \\ \text{AGENTIVE} = \ldots & \end{array}\right] \end{array}\right]$$

The analysis of AGENTIVE modification is also more complex. In addition to *di*, *della* is also found for subtyping of arguments in the AGENTIVE. In other cases, such as *morte da annegamento*, *death from drowning* and *bruciatura da sole*, *sun burn*, the preposition is *da*. This preposition *da* has a different meaning from the one associated with the TELIC. It corresponds to the English preposition *from* and it is interpreted as introducing an experiencing relation. It is found in cases where the head noun is an event and the modifier introduces the causal factor which brought about that event. We turn now to consider some of the applications of this work in more detail.

7 Applications

The analysis of complex nominal constructions presented in this paper has a range of important applications in natural language processing. Complex nominals play an important role in the encapsulation and expression of nominal concepts and are frequent in a wide variety of types of texts. Therefore, the ability to handle complex nominals is essential for parsing and generation systems for either English or Italian. It is important to note that systems utilizing compositional apparatus for the analysis of complex nominals need not treat all compounds compositionally. The optimal arrangement will be to list frequent and idiosyncratic compound forms in the lexicon and use the compositional apparatus for forms which are not listed, or in instances when the listed interpretation is ruled out by context. We would also like to point out that we do not expect to develop an analysis which will handle all and every compound form. Our target is to have an account which will handle the majority of productive compounding patterns. Another important use of the compositional apparatus described here is in lexical acquisition of compound forms. This machinery can be used to indicate potential interpretations for compounds. A human editor can then select the appropriate interpretation from the candidate set and add have the compound added to the lexicon.

Given the range of different semantic relations that can hold between the elements of a complex nominal, they are frequently ambiguous. English compounds are worse than Italian post-modified forms in this respect, since in Italian the preposition gives at least some indication of the relation involved. The approach described in this paper constrains the interpretation of complex nominals using the type system. For example, the schema in (6), which accounts for *bread knife*, requires the modifying noun to be typed as **individual**. This limits the set of potential modifiers to those typed as **individual**. Since the content of the modifier is structure-shared with an argument position within the TELIC, this set of potential modifiers is further constrained by type constraints imposed by the relation in the TELIC role. The *cut_act* will require the object cut to be a separable object. It could potentially require the cutter to be significantly harder than the object to be cut. Type constraints of this kind serve to greatly reduce the degree of ambiguity in a given complex nominal, but it will still generally be the case that more than one interpretation is predicted for a given form. For example, a form like *bone knife* could be interpreted either as a knife used for cutting bone or a knife made of bone. The approach described here needs to be integrated with further mechanisms and heuristics in order to determine the best guess for complex nominal interpretation in any given case. One important class of mechanisms are those which examine the current sentential and discourse context in order to restrict the range of interpretations. For example, if *bone knife* appears in a medical text, *bone* most probably specifies the object to be cut by the knife, while if it shows up in a text concerning prehistoric man, *bone* most probably refers to the constitution of the knife. One way in which compounds can be further disambiguated is through the incorporation of a statistical model as one of the heuristics employed in determining the appropriate interpretation. In such an approach, one could train on a data set comprised of compounds paired with an indication of the relation holding between the head and the modifier. The resulting model would provide the probability that a given complex nominal involves a particular kind of modification relation. In order to have useful predictive power, it would be best to assign semantic types to the elements of the complex nominal and determine the probability that a complex nominal consisting of words of types A and B involves modification relation C. Given the sparsity of data to support a statistically based approach we believe that the way forward in this area is to pursue the integration of a rule-based approach with a statistical model. Such integration has already proven effective in the treatment of sense extension phenomena (Copestake and Briscoe 1995). We leave further investigation of this integration for future work.

This work also has important consequences for applications in multilingual natural language processing. The most obvious of these is the use of a cross-linguistic approach to complex nominals in machine translation. Translation of complex nominals from Italian to English will be more straightforward, since there is a loss of information rather than a gain. It is important to note, however, that not all Italian complex nominals involving post-modification can be translated as noun-noun compounds in English. For example, forms such as *coltello* *da* *macellaio* (literally, knife of butcher), in which the modifier is an agent using the object described by the head, does not translate as ***butcher knife***. In English, the appropriate nominal construction in this case uses the possessive: ***butcher's knife***.

Translation from English to Italian is substantially more difficult given the difference in explicitness regarding the semantic relation between the head and modifier. In order to generate the proper output in Italian, it is necessary to determine the relation between the elements in the English compound structure and to determine the appropriate preposition in Italian for expression of that relation. One approach to this task is to use the GL representation language essentially as an interlingua (McDonald 1995). The phrase structure schemata for English are used in order to determine potential interpretations for a given English compound construction. The most likely interpretation from the candidate set is picked on the basis of contextual and statistical models. The CONTENT of the chosen candidate is then matched against the outputs of the various phrase structure schemata used for Italian. When an appropriate schema is identified it is instantiated with lexical items from the Italian lexicon in order to generate the Italian translation. An important feature of this approach is that it utilizes resources which are independently needed for analysis of the languages involved. Aside from translation, the phrase structure schemata can also be used for multi-lingual generation. If a particular concept is encoded in the GL lexical representation language, the language-specific phrase structure schemata can be employed to generate the corresponding complex nominal in each language.

In addition to the importance of successful translation of complex nominals for full-text machine translation, this functionality is useful in itself for applications in multi-lingual information retrieval and information extraction. Since complex nominals are so frequently used to coin terms which encapsulate important distinguished concepts within a domain, their successful identification and processing is an essential element of determination of the topic of a text and they provide important hooks for information retrieval. In a multi-lingual setting, such as information retrieval over the World Wide Web, it may be desirable for a search for a complex nominal from one language to yield documents regarding the same concept in other languages. The approach to translation of complex nominals described above enables this functionality. For a given form compound form in English it is possible to determine potential realizations of that form in Italian.

8 Conclusion

In this paper, we have shown how the theory of qualia structure within the Generative Lexicon, enables a compositional treatment of compounds. In compounds where the modifying noun describes an individual, in composition, the modifier further specifies the type of an argument to a predicate in the TELIC, AGENTIVE, or CONSTITUTIVE role. In Italian, the canonical prepositions for these three kinds of modification are *da*, *di*, and *a*, respectively. In compounds where the modifying noun denotes an event, the composition in the compound frequently involves co-composition between the qualia structure of the head and modifier. In Italian, for TELIC modification the preposition is *da* when the modifier describes an activity and *di* when the modifier describes a result. In addition to its theoretical relevance, the approach to the semantics of complex nominals described here has important applications in the construction of natural language processing systems. In particular, it provides the foundations for machine translation of complex nominals between English and Italian and can be readily applied in multi-lingual generation and multi-lingual information extraction.

References

Alshawi, Hiyan.1987. *Memory and Context for Language Interpretation.* Studies in Natural Language Processing. Cambridge University Press, Cambridge, England.

Beard, Robert. 1996. *Head Operations and Head-Modifier Ordering in Nominal Compounds.* Presentation at 1996 Linguistic Society of America Meeting, San Diego, California.

Bergsten, N. 1991. *A Study on Compound Substantives in English.* Almquist and Wiksell, Uppsala.

Bouillon, Pierette. 1995. *The Semantics of Adjectival Modification.* ms. ISSCO, Geneva.

Bouillon, P, K. Bösefeldt, and Graham Russell. 1992. Compound Nouns in a Unification-Based MT System. In *Proceedings of the Third Conference on Applied Natural Language Processing* (p209-215). Trento, Italy.

Busa, Federica. 1996. *Compositionality and the Semantics of Nominals.* Doctoral Dissertation. Brandeis University.

Copestake, Ann., and Ted Briscoe. 1995. Semi-productive Polysemy and Sense Extension. Journal of Semantics 12.

Downing, P. 1977. On the Creation and Use of English Compound Nouns. Language 53. 810-842.

Finin, Timothy. W. 1980. *The Semantic Interpretation of Compound Nominals.* Doctoral Dissertation. University of Illinois at Urbana-Champaign.

Flickinger, Daniel. 1987. Lexical Rules in the Hierarchical Lexicon. Doctoral Dissertation. Stanford University.

Hobbs, Jerry R., Martin. E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as Abduction. In Fernando C.N. Pereira and Barbara Grosz (eds.) *Natural Language Processing.* MIT Press, Cambridge, Massachusetts.

Isabelle, P. 1984. Another Look at Nominal Compounds. In *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Meeting of the ACL.* (pp. 509-516).

Jespersen, Otto. 1942. *A Modern English Grammar on Historical Principles, IV.* Munksgaard, Copenhagen.

Jones, Bernard. 1995. Nominal Compounds and Lexical Rules. *Working Notes of the Acquilex Workshop on Lexical Rules.* Cambridge, England, August 1995.

Johnston, Michael, Branimir Boguraev, and James Pustejovsky. 1995. The Acquisition and Interpretation of Complex Nominals. *Working Notes of AAAI Spring Symposium on the Representation and Acquisition of Lexical Knowledge*, Stanford University, Palo Alto, California.

Lees, Robert. 1970. Problems in the Grammatical Analysis of English Nominal Compounds. In Bierwisch and Heidolph (eds.) *Progress in Linguistics.* Mouton, The Hague.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, New York.

Marchand, Hans. 1969. *The Categories and Types of Present Day English Word Formation.* C.H Becksche, Munich.

McDonald, David. 1995. *Lexical Discontinuities in the Functional Meaning of Words.* Working Notes of Multilingual Text Generation Workshop. IJCAI, August 20-21, Montreal, Quebec.

McDonald, David B. 1982. *Understanding Noun Compounds.* CMU Technical Report CS-82-102.

Pollard, Carl and Ivan Sag. 1987. Information-based Syntax and Semantics, Volume 1: Fundamentals. CSLI Lecture Notes Series No.13. Centre for the Study of Language and Information. Stanford University.

Pollard, Carl and Ivan Sag. 1994. Head-driven Phrase Structure Grammar. University of Chicago Press. Chicago.

Pustejovsky, James. 1991. The Generative Lexicon. *Computational Linguistics.* 17.4.

Pustejovsky, James. 1995. *The Generative Lexicon.* MIT Press, Cambridge, Massachusetts.

Warren, Beatrice, 1987. *Semantic Patterns of Noun-Noun Compounds.* Gothenburg Studies in English 41. Acta Universitatis Gothoburgensis, Gothenburg.