The Computational Lexical Semantics of Syntagmatic Relations

Evelyne Viegas, Stephen Beale and Sergei Nirenburg
New Mexico State University
Computing Research Lab,
Las Cruces, NM 88003,
USA
viegas,sb,sergei@crl.nmsu.edu

Abstract

In this paper, we address the issue of syntagmatic expressions from a computational lexical semantic perspective. From a representational viewpoint, we argue for a hybrid approach combining linguistic and conceptual paradigms, in order to account for the continuum we find in natural languages from free combining words to frozen expressions. In particular, we focus on the place of lexical and semantic restricted co-occurrences. From a processing viewpoint, we show how to generate/analyze syntagmatic expressions by using an efficient constraint-based processor, well fitted for a knowledge-driven approach.

1 Introduction

You can take advantage of the chambermaid[1] is not a collocation one would like to generate in the context of a hotel to mean "use the services of." This is why collocations should constitute an important part in the design of Machine Translation or Multilingual Generation systems.

In this paper, we address the issue of syntagmatic expressions from a computational lexical semantic perspective. From a representational viewpoint, we argue for a hybrid approach combining linguistic and conceptual paradigms, in order to account for the continuum we find in natural languages from free combining words to frozen expressions (such as in idioms *kick the (proverbial) bucket*). In particular, we focus on the representation of restricted semantic and lexical co-occurrences, such as *heavy smoker* and *professor ... students* respectively, that we define later. From a processing viewpoint, we show how to generate/analyze syntagmatic expressions by using an efficient constraint-based processor, well fitted for a knowledge-driven approach. In the following, we first compare different approaches to collocations. Second, we present our approach in terms of representation and processing. Finally, we show how to facilitate the acquisition of co-occurrences by using 1) the formalism of lexical rules (LRs), 2) an inheritance hierarchy of Lexical Semantic Functions (LSFs).

[1] Lederer, R. 1990. *Anguished English* A Laurel Book, Dell Publishing.

2 Approaches to Syntagmatic Relations

Syntagmatic relations, also known as collocations, are used differently by lexicographers, linguists and statisticians denoting almost similar but not identical classes of expressions.

The traditional approach to collocations has been **lexicographic**. Here dictionaries provide information about what is unpredictable or idiosyncratic. Benson (1989) synthesizes Hausmann's studies on collocations, calling expressions such as *commit murder, compile a dictionary, inflict a wound, etc.* "fixed combinations, recurrent combinations" or "collocations". In Hausmann's terms (1979) a collocation is composed of two elements, a *base* ("Basis") and a *collocate* ("Kollokator"); the base is semantically autonomous whereas the collocate cannot be semantically interpreted in isolation. In other words, the set of lexical collocates which can combine with a given basis is not predictable and therefore collocations must be listed in dictionaries.

It is hard to say that there has been a real focus on collocations from a **linguistic** perspective. The lexicon has been broadly sacrificed by both English-speaking schools and continental European schools. The scientific agenda of the former has been largely dominated by syntactic issues until recently, whereas the latter was more concerned with pragmatic aspects of natural languages. The focus has been on grammatical collocations such as *adapt to, aim at, look for.* Lakoff (1970) distinguishes a class of expressions which cannot undergo certain operations, such as nominalization, causativization: *the problem is hard; *the hardness of the problem; *the problem hardened.* The restriction on the application of certain syntactic operations can help define collocations such as *hard problem*, for example. Mel'čuk's treatment of collocations will be detailed below.

In recent years, there has been a resurgence of **statistical** approaches applied to the study of natural languages. Sinclair (1991) states that *"a word*

which occurs in close proximity to a word under investigation is called a collocate of it. ... Collocation is the occurrence of two or more words within a short space of each other in a text". The problem is that with such a definition of collocations, even when improved,[2] one identifies not only collocations but free-combining pairs frequently appearing together such as *lawyer-client; doctor-hospital.* However, nowadays, researchers seem to agree that combining statistic with symbolic approaches lead to quantifiable improvements (Klavans and Resnik, 1996).

The Meaning Text Theory Approach The Meaning Text Theory (MTT) is a generator-oriented lexical grammatical formalism. Lexical knowledge is encoded in an entry of the Explanatory Combinatorial Dictionary (ECD), each entry being divided into three zones: the *semantic zone* (a semantic network representing the meaning of the entry in terms of more primitive words), the *syntactic zone* (the grammatical properties of the entry) and the *lexical combinatorics zone* (containing the values of the **Lexical Functions** (LFs) [3]). LFs are central to the study of collocations:

> *A lexical function F is a correspondence which associates a lexical item L, called the key word of F, with a set of lexical items F(L)-the value of F. (Mel'čuk, 1988)*[4]

We focus here on *syntagmatic* LFs describing co-occurrence relations such as *pay attention, legitimate complaint; from a distance.*[5]

Heylen et al. (1993) have worked out some cases which help license a starting point for assigning LFs. They distinguish four types of syntagmatic LFs:

- *evaluative qualifier*
  Magn(bleed) = profusely
- *distributional qualifier*
  Mult(sheep) = flock
- *co-occurrence*
  Loc-in(distance)= at a distance
- *verbal operator*
  Oper1(attention) = pay

The MTT approach is very interesting as it provides a model of production well suited for generation with its different strata and also a lot of lexical-semantic information. It seems nevertheless that all the collocational information is listed in a static way. We believe that one of the main drawbacks of the approach is the lack of any predictable calculi on the possible expressions which can collocate with each other **semantically.**

3 The Computational Lexical Semantic Approach

In order to account for the continuum we find in natural languages, we argue for a continuum perspective, spanning the range from free-combining words to idioms, with semantic collocations and idiosyncrasies in between as defined in (Viegas and Bouillon, 1994):

- **free-combining words** (*the girl ate candies*)
- **semantic collocations** (*fast car; long book*)[6]
- **idiosyncrasies** (*large coke; green jealousy*)
- **idioms** (*to kick the (proverbial) bucket*)

Formally, we go from a purely compositional approach in "free-combining words" to a non-compositional approach in idioms. In between, a (semi-)compositional approach is still possible. (Viegas and Bouillon, 1994) showed that we can reduce the set of what are conventionally considered as idiosyncrasies by differentiating "true" idiosyncrasies (difficult to derive or calculate) from expressions which have well-defined calculi, being compositional in nature, and that have been called semantic collocations. In this paper, we further distinguish their idiosyncrasies into:

- **restricted semantic co-occurrence**, where the meaning of the co-occurrence is semi-compositional between the base and the collocate (*strong coffee, pay attention, heavy smoker, ...*)
- **restricted lexical co-occurrence**, where the meaning of the collocate is compositional but has a lexical idiosyncratic behavior (*lecture ... student; rancid butter; sour milk*).

We provide below examples of restricted semantic co-occurrences in (1), and restricted lexical co-occurrences in (2).

Restricted semantic co-occurrence The semantics of the combination of the entries is semi-compositional. In other words, there is an entry in the lexicon for the base, (the semantic collocate is encoded inside the base), whereas we cannot directly refer to the sense of the semantic collocate in the lexicon, as it is not part of its senses. We assign the co-occurrence a new semi-compositional sense,

[2] Church and Hanks (1989), Smadja (1993) use statistics in their algorithms to extract collocations from texts.

[3] See (Iordanskaja et al., 1991) and (Ramos et al., 1994) for their use of LFs in MTT and NLG respectively.

[4] (Heid, 1989) contrasts Hausman's base and collate to Mel'čuk's keyword and LF values.

[5] There are about 60 LFs listed said to be universal; the lexicographic approach of Mel'čuk and Žolkovsky has been applied among other languages to Russian, French, German and English.

[6] See (Pustejovsky, 1995) for his account of such expressions using a coercion operator.

where the sense of the base is composed with a new sense for the collocate.

```
(1a) #0=[key: "smoker",
         rel: [syntagmatic: LSFIntensity
               [base: #0, collocate:
                [key: "heavy",
                 gram: [subCat: Attributive,
                        freq: [value: 8]]]]] ...]

(1b) #0=[key: "attention",
         rel: [syntagmatic: LSFOper
               [base: #0, collocate:
               [key: "pay",
                gram: [subCat: SupportVerb,
                       freq: [value: 5]]]]] ...]
```

In examples (1), the LSFs (LSFIntensity, LSFOper, ...) are equivalent (and some identical) to the LFs provided in the ECD. The notion of LSF is the same as that of LFs. However, LSFs and LFs are different in two ways: i) conceptually, LSFs are organized into an inheritance hierarchy; ii) formally, they are rules, and produce a new entry composed of two entries, the base with the collocate. As such, the new composed entry is ready for processing. These LSFs signal a compositional syntax and a semi-compositional semantics. For instance, in (1a), a *heavy smoker* is somebody who smokes a lot, and not a "fat" person. It has been shown that one cannot code in the lexicon all uses of *heavy* for *heavy smoker, heavy drinker, ...* Therefore, we do not have in our lexicon for *heavy* a sense for "a lot", or a sense for "strong" to be composed with *wine*, etc... It is well known that such co-occurrences are lexically marked; if we allowed in our lexicons a proliferation of senses, multiplying ambiguities in analysis and choices in generation, then there would be no limit to what could be combined and we could end up generating **heavy coffee* with the sense of "strong" for *heavy*, in our lexicon.

The left hand-side of the rule LSFIntensity specifies an "Intensity-Attribute" applied to an event which accepts aspectual features of duration. In (1a), the event is *smoke*. The LSFIntensity also provides the syntax-semantic interface, allowing for an Adj-Noun construction to be either predicative (*the car is red*) or attributive (*the red car*). We need therefore to restrict the co-occurrence to the Attributive use only, as the predicative use is not allowed: (*the smoker is heavy*) has a literal meaning or figurative, but not collocational.

In (1b) again, there is no sense in the dictionary for *pay* which would mean *concentrate*. The rule LSFOper makes the verb a verbal operator. No further restriction is required.

Restricted lexical co-occurrence The semantics of the combination of the entries is compositional. In other words, there are entries in the lexicon for the base and the collocate, with the same senses as in the co-occurrence. Therefore, we can directly refer to the senses of the co-occurring words. What we are capturing here is a lexical idiosyncrasy or in other words, we specify that we should prefer this particular combination of words. This is useful for analysis, where it can help disambiguate a sense, and is most relevant for generation; it can be viewed as a preference among the paradigmatic family of the co-occurrence.

```
(2a) #0=[key: "truth",
         rel: [syntagmatic: LSFSyn
               [base: #0, collocate:
               [key: "plain", sense: adj2,
                lr: [comp:no, superl:no]]]] ...]

(2b) #0=[key: "pupil",
         rel: [syntagmatic: LSFSyn
               [base: #0, collocate:
               [key: "teacher", sense: n2,
                freq: [value: 5]]]]...]

(2c) #0=[key: "conference",
        rel: [syntagmatic: LSFSyn
              [base: #0, collocate:
              [key: "student", sense: n1,
               freq: [value: 9]]]] ...]
```

In examples (2), the LSFSyn produces a new entry composed of two or more entries. As such, the new entry is ready for processing. LSFSyn signals a compositional syntax and a compositional semantics, and restricts the use of lexemes to be used in the composition. We can directly refer to the sense of the collocate, as it is part of the lexicon.

In (2a) the entry for *truth* specifies one co-occurrence (*plain truth*), where the sense of *plain* here is adj2 (obvious), and not say adj3 (flat). The syntagmatic expression inherits all the zones of the entry for "plain", sense adj2, we only code here the irregularities. For instance, "plain" can be used as "plainer" "plainest" in its "plain" sense in its adj2 entry, but not as such within the lexical co-occurrence "*plainer truth", "*plainest truth", we therefore must block it in the collocate, as expressed in (comp: no, superl: no). In other words, we will not generate "plainer/plainest truth". Examples (2b) and (2c) illustrate complex entries as there is no direct grammatical dependency between the base and the collocate. In (2b) for instance, we prefer to associate *teacher* in the context of a *pupil* rather than any other element belonging to the paradigmatic family of *teacher* such as *professor, instructor.*

Formally, there is no difference between the two types of co-occurrences. In both cases, we specify the base (which is the word described in the en-

try itself), the collocate, the frequency of the co-occurrence in some corpus, and the LSF which links the base with the collocate. Using the formalism of typed feature structures, both cases are of type Co-occurrence as defined below:

```
Co-occurrence = [base: Entry,
                 collocate: Entry,
                 freq: Frequency];
```

3.1 Processing of Syntagmatic Relations

We utilize an efficient constraint-based control mechanism called *Hunter-Gatherer* (HG) (Beale, 1997). HG allows us to mark certain compositions as being dependent on each other and then forget about them. Thus, once we have two lexicon entries that we know go together, HG will ensure that they do. HG also gives preference to co-occurring compositions. In analysis, meaning representations constructed using co-occurrences are preferred over those that are not, and, in generation, realizations involving co-occurrences are preferred over equally correct, but non-cooccurring realizations.[7]

The real work in processing is making sure that we have the correct two entries to put together. In restricted semantic co-occurrences, the co-occurrence does not have the correct sense in the lexicon. For example, when the phrase *heavy smoker* is encountered, the lexicon entry for *heavy* would not contain the correct sense. (1a) could be used to create the correct entry. In (1a), the entry for *smoker* contains the key, or trigger, *heavy*. This signals the analyzer to produce another sense for *heavy smoker*. This sense will contain the same syntactic information present in the "old" *heavy*, except for any modifications listed in the "gram" section (see (1a)). The semantics of the new sense comes directly from the LSF. Generation works the same, except the trigger is different. The input to generation will be a SMOKE event along with an Intensity-Attribute. (1a), which would be used to realize the SMOKE event, would trigger LSFIntensify which has the Intensity-Attribute in the left hand-side, thus confirming the production of *heavy*.

Restricted lexical co-occurrences are easier in the sense that the correct entry already exists in the lexicon. The analyzer/generator simply needs to detect the co-occurrence and add the constraint that the corresponding senses be used together. In examples like (2b), there is no direct grammatical or semantic relationship between the words that co-occur. Thus, the entire clause, sentence or even text may have to be searched for the co-occurrence. In practice, we limit such searches to the sentence level.

[7]The selection of co-occurrences is part of the lexical process, in other words, if there are reasons not to choose a co-occurrence because of the presence of modifiers or because of stylistics reasons, the generator will not generate the co-occurrence.

3.2 Acquisition of Syntagmatic Relations

The acquisition of syntagmatic relations is knowledge intensive as it requires human intervention. In order to minimize this cost we rely on conceptual tools such as lexical rules, on the LSF inheritance hierarchy.

Lexical Rules in Acquisition The acquisition of restricted semantic co-occurrences can be minimized by detecting rules between different classes of co-occurrences (modulo presence of derived forms in the lexicon with same or subsumed semantics). Looking at the following example,

A	+	N	<=>	V	+	Adv
bitter		resentment		resent		bitterly
heavy		smoker		smoke		heavily
big		eater		eat		*bigly

V	+	Adv	<=>	Adv	+	Adj-ed
oppose		strongly		strongly		opposed
oblige		morally		morally		obliged

we see that after having acquired with human intervention co-occurrences belonging to the A + N class, we can use lexical rules to derive the V + Adv class and also Adv + Adj-ed class.

Lexical rules are a useful conceptual tool to extend a dictionary. (Viegas et al., 1996) used derivational lexical rules to extend a Spanish lexicon. We apply their approach to the production of restricted semantic co-occurrences. Note that *eat bigly* will be produced but then rejected, as the form *bigly* does not exist in a dictionary. The rules overgenerate co-occurrences. This is a minor problem for analysis than for generation. To use these derived restricted co-occurrences in generation, the output of the lexical rule processor must be checked. This can be done in different ways: dictionary check, corpus check and ultimately human check.

Other classes, such as the ones below can be extracted using lexico-statistical tools, such as in (Smadja, 1993), and then checked by a human.

```
V + N     pay attention, meet an obligation,
          commit an offence, ...

N + N     dance marathon, marriage ceremony
          object of derision, ...
```

LSFs and Inheritance We take advantage of 1) the semantics encoded in the lexemes, and 2) an inheritance hierarchy of LSFs. We illustrate briefly this notion of LSF inheritance hierarchy. For instance, the left hand-side of LSFChangeState specifies that it applies to foods (solid or liquid) which are human processed, and produces the collocates *rancid*, *rancio* (Spanish). Therefore it could apply to *milk*, *butter*, or *wine*. The rule would end up

producing *rancid milk*, *rancid butter*, or *vino rancio* (rancid wine) which is fine in Spanish. We therefore need to further distinguish LSFChangeState into LSFChangeStateSolid and LSFChangeStateLiquid. This restricts the application of the rule to produce *rancid butter*, by going down the hierarchy. This enables us to factor out information common to several entries, and can be applied to both types of co-occurrences. We only have to code in the co-occurrence information relevant to the combination, the rest is inherited from its entry in the dictionary.

4 Conclusion

In this paper, we built on a continuum perspective, knowledge-based, spanning the range from free-combining words to idioms. We further distinguished the notion of idiosyncrasies as defined in (Viegas and Bouillon, 1994), into restricted semantic co-occurrences and restricted lexical co-occurrences. We showed that they were formally equivalent, thus facilitating the processing of strictly compositional and semi-compositional expressions. Moreover, by considering the information in the lexicon as constraints, the linguistic difference between compositionality and semi-compositionality becomes a virtual difference for Hunter-Gatherer. We showed ways of minimizing the acquisition costs, by 1) using lexical rules as a way of expanding co-occurrences, 2) taking advantage of the LSF inheritance hierarchy. The main advantage of our approach over the ECD approach is to use the semantics coded in the lexemes along with the language independent LSF inheritance hierarchy to propagate restricted semantic co-occurrences. The work presented here is complete concerning representational aspects and processing aspects (analysis and generation): it has been tested on the translations of on-line unrestricted texts. The large-scale acquisition of restricted co-occurrences is in progress.

5 Acknowledgements

This work has been supported in part by DoD under contract number MDA-904-92-C-5189. We would like to thank Pierrette Bouillon, Léo Wanner and Rémi Zajac for helpful discussions and the anonymous reviewers for their useful comments.

References

S. Beale. 1997. *HUNTER-GATHERER: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to Computational Semantics.* Ph.D. Diss., Carnegie Mellon University.

M. Benson. 1989. The Structure of the Collocational Dictionary. In *International Journal of Lexicography.*

K.W. Church and P. Hanks. 1989. Word Association Norms, Mutual Information and Lexicography. In Proceedings of *the 27th Annual Meeting of the Association for Computational Linguistics.*

F.J. Hausmann. 1979. Un dictionnaire des collocations est-il possible ? In *Travaux de Linguistique et de Littérature XVII, 1.*

U. Heid. 1979. *Décrire les collocations : deux approches lexicographiques et leur application dans un outil informatisé.* Internal Report, Stuttgart University.

D. Heylen. 1993. Collocations and the Lexicalisation of Semantic Information. In *Collocations*, TR ET-10/75, Taaltechnologie, Utrecht.

L. Iordanskaja, R. Kittredge and A. Polguère. 1991. Lexical Selection and Paraphrase in a Meaning-text Generation Model. In C. L. Paris, W. Swartout and W. Mann (eds), *NLG in AI and CL.* Kluwer Academic Publishers.

J. Klavans and P. Resnik. 1996. *The Balancing Act, Combining Symbolic and Statistical Approaches to Language.* MIT Press, Cambridge Mass., London England.

G. Lakoff. 1970. *Irregularities in Syntax.* New York: Holt, Rinehart and Winston, Inc.

I. Mel'čuk. 1988. Paraphrase et lexique dans la théorie Sens-Texte. In Bes & Fuchs (ed) *Lexique6.*

S. Nirenburg and I. Nirenburg. 1988. A Framework for Lexical Selection in NLG. In Proceedings of *COLING 88.*

J. Pustejovsky. 1995. *The Generative Lexicon.* MIT Press.

M. Ramos, A. Tutin and G. Lapalme. 1994. Lexical Functions of Explanatory Combinatorial Dictionary for Lexicalization in Text Generation. In P. St-Dizier & E. Viegas (Ed) *Computational Lexical Semantics*: CUP.

J. Sinclair. 1991. *Corpus, Concordance, Collocations.* Oxford University Press.

F. Smadja. 1993. Retrieving Collocations from Texts: Xtract. *Computational Linguistics, 19(1).*

E. Viegas and P. Bouillon. 1994. Semantic Lexicons: the Cornerstone for Lexical Choice in Natural Language Generation. In Proceedings of the *7th INLG*, Kennebunkport.

E. Viegas, B. Onyshkevych, V. Raskin and S. Nirenburg. 1996. From *Submit* to *Submitted* via *Submission*: on Lexical Rules in Large-scale Lexicon Acquisition. In Proceedings of *the 34th Annual Meeting of the Association for Computational Linguists.*

L. Wanner. 1996. *Lexical Functions in Lexicography and Natural Language Processing.* John Benjamin Publishing Company.