Completing Mixed Language Grammars Through Womb Grammars Plus Ontologies*

Ife Adebara[1], Veronica Dahl[1,3] and Sergio Tessaris[2]
[1]*Department of Computer Science, Simon Fraser University, 8888 University Drive, Burnaby, Canada*
[2]*Faculty of Computer Science, Free University of Bozen–Bolzano, Piazza Domenicani 3, 39100 Bolzano, Italy*
[3]*Institute of Software Engineering and Compiler Construction, University of Ulm, Ulm, Germany*

Keywords: Self-modifying Grammars, Womb Grammars, Ontologies, Mixed Language Text, Partially Known Grammars, Constraint Acquisition, Universal Language, Parsing, CHRG (Constraint Handling Rule Grammars), Constraint Based Grammars, Property Grammars.

Abstract: Womb Grammars are a recently introduced constraint-based methodology for acquiring linguistic information on a given language from that of another, implemented in CHRG (Constraint Handling Rule Grammars). This is a position paper that discusses their possible adaptation to multilingual text parsing. In particular, we propose to detect unspecified information with appropriate ontologies. Our proposed methodology exploits the descriptive power of constraints both for defining sentence acceptability and for inferring lexical knowledge from a word's sentential context, even when foreign.

1 INTRODUCTION

Social media promotes communication across countries, multiplying the opportunities for users to spontaneously mix syntax, lexicons and jargons. Also, there are domains where syntactic arrangements different from the standard arrangement are acceptable. These factors, together with the increasing infiltration of English words and specific group jargons into technical and even every day communications in many other languages, results in the need for ever more flexible parsers if we are to succeed in extracting information from text in timely fashion. Yet we are quite far from being able to address the challenges inherent in multilingual and creative text. In fact, one of the worst nightmares for linguistics is that of trying to parse textual sources that do not respect the standard grammar.

Traditional parsers focus on constructing syntactic trees for complete and correct sentences in a given language. More flexible parsing models can be arrived at in economic fashion by giving up syntactic trees as a focus and focusing instead on *grammar constraints*, also called *properties*. For instance, if we were to work with tree-oriented rules such as:

```
np --> det, adj, n.
```

their adaptation into a language where nouns must precede adjectives would require changing every rule where these two constituents are involved. In contrast, by expressing the same rule in terms of separate constraints, we only need to change the precedence constraint into saying that adjectives must precede nouns, and the modification carries over to the entire grammar without further ado.

In this paper we propose to combine Womb Grammar parsing- a property-based methodology for multilingual parsing developed by Dahl and Miralles (Dahl and Miralles, 2012)- with ontologies, in view of further specifying partial information which can be lexical or structural, in an automatic manner.

2 BACKGROUND

Womb Grammars (Dahl and Miralles, 2012) were designed for inducing a target language's syntax from the known syntax of a source language plus a representative corpus of correct sentences in the target language. As such they can be considered a kind of self-modifying grammar, whose approach is quite different from that of predecessors (e.g. (Jackson, 2006)

*This research was supported by NSERC Discovery grant 31611024 and was started during a visit by Veronica Dahl and Sergio Tessaris to Universidade Nova de Lisboa.

Adebara I., Dahl V. and Tessaris S..
Completing Mixed Language Grammars Through Womb Grammars Plus Ontologies.
DOI: 10.5220/0005353202920297
In *Proceedings of the International Conference on Agents and Artificial Intelligence* (PUaNLP-2015), pages 292-297
ISBN: 978-989-758-073-4

resorts heavily to push-down automata; (Christiansen, 2011), while being more declarative, are an extension of attribute grammars. Womb grammars, in contrast, are constraint-based: they derive a target language's syntax by observing the list of violated properties that are output when correct sentences in the target language are fed to the source grammar, and correcting that grammar so that these properties are no longer violated.

In the original Womb Grammar formalism, we had two languages: the source language, of which both the syntax and the lexicon were known, and the target language, of which only the lexicon and a correct input corpus were known. Here we still assume a main language such as English, but it might be creatively cross fertilised with multilingual contributions, both in structure and lexicon, from other languages.

3 OUR PROPOSED METHODOLOGY

The main difficulty in adapting our methodology is that the target language's input can no longer be considered correct. We shall first consider lexical and structural intrusions separately, and then discuss how to deal with them jointly.

Before doing so, let us briefly recall how lexical items are recognized and how constraints are enforced by the Womb Grammar parser.

3.1 A Few Implementation Details

Our implementation of Womb Grammars (Dahl and Miralles, 2012) is done in terms of CHRG, or Constraint handling Rule Grammars (Christiansen, 2005). Below we show some actual code for completeness, but our description should be intuitively clear enough for those readers with no background on CHR to follow.

Each word is stored in a CHRG symbol word/3, along with its category and traits (i.e. word(n,[sing,masc],livre)).

Grammar constraints are entered in terms of a CHRG constraint g/1, whose argument stores each possible grammar property. For instance, an English noun phrase parser would include the constraints: g(obligatority(n)), g(constituency(det)), g(precedence(det,adj)), g(unicity(det)), g(requirement(n,det)), g(dependence(det,n)), and so on. These properties are weeded out upon detection of a violation by CHRG rules that look for them, e.g. an input noun phrase where an adjective precedes a noun will provoke deletion of the constraint g(precedence(n,adj)) plus perhaps (if the rest of the input corpus warrants it) inclusion of the converse constraint: g(precedence(adj,n)). The following CHRG rule accomplishes that:

```
!word(C2,_,_), ... , !word(C1,_,_),
  {g(precedence(C1,C2))} <:>
  {update(precedence(C1,C2))}.
```

Note that the rule works bottom-up, and that the three dots are a facility of CHRG which allows us to skip over an unspecified substring of words. The curly brackets indicate a call to a procedure (as opposed to a grammar symbol).

The CHRG parse predicate stores and abstracts the position of each word in the sentence. In plain English, the above rule states that if a word of category C2 precedes a word of category C1, and there is a precedence rule stipulating that words of category C1 must precede words of category C2, the precedence-updating rule needs to be invoked (in CHRG syntax the symbols prefixed with exclamation points are kept, while the ones without are replaced by the body of the rule, in this case an update constraint that invokes some housekeeping procedures).

Each of the properties dealt with has similar rules associated with it.

3.2 Underspecified Lexical Categories

Let us first consider the problem of accommodating extraneous words. We assume in a first stage that we have only one language with known syntax and lexicon, and an input corpus which is correct save for the occasional intrusion of neologisms or words belonging to another language or jargon. We can adapt our Womb Grammar methodology to this situation, by running the input corpus as is and observing the list of violated properties that will be output. Since we know everything to be correct except that some lexical items do not "belong", we know that the violated properties stem from those lexical items that failed to parse. By examining the violated properties, we can draw useful inferences about the lexical items in question. For instance, if the head noun appears as an unknown word, among the violated properties we will read that the obligatory character of a noun phrase's noun has been violated, which can lead us to postulate that the word in question is a noun. A violated exigency property would likewise suggest that the unrecognised word has the category that is required and has not been found.

It is clear that with sufficient programming effort, any computational linguistic methodology can

be adapted to guess lexical categories of extraneous words from context. However in most of them, this would require a major modification of the parser. Take for instance DCGs (Definite Clause Grammars, (Pereira and Warren, 1980)), where lexical rules would appear as exemplified by:

```
noun --> [borogove].
```

If the lexicon does not explicitly include the word "borogrove" among the nouns, the parser would simply fail when encountering it. One could admit unknown nouns through the following rule:

```
noun --> [_].
```

But since this rule would indiscriminately accept any word as a noun (and similar rules would have to be included in order to treat possible extraneous words in any other category), this approach would mislead the parser into trying countless paths that are doomed to fail, and might even generate wrong results.

In contrast, we can parse extraneous words through Womb Grammar by anonymizing the category and its features rather than the word itself, e.g. word(Category,[Number,Gender],borogrove)), which more accurately represents what we know and what we don't. The category and features will become efficiently instantiated through constraint satisfaction, taking into account all the properties that must be satisfied by this word in interaction with its context.

Of course, what would be most interesting would be to derive the meaning of the word that "does not belong". While Womb Grammars do not yet have a complete way of treating semantics, the clues they can provide regarding syntactic category can serve to guide a subsequent semantic analysis, or to bypass the need for a complete semantic analysis by the concomitant use of ontologies relevant to domain-specific uses of our parser. In general, we are not necessarily interested in capturing the exact meaning of each unrecognised word; but rather to infer its relation with known words. The problem can be casted into the (automatic) extraction of a portion of the hypernym relation involving the extraneous word using the actual document or additional sources as corpora (see (Clark et al., 2012)).

For instance, in the poem "Jabberwocky", by Lewis Carroll, nonsense words are interspersed within English text with correct syntax. Our target lexicon, which we might call Wonderland Lexicon or WL, can be to some extent reconstructed from the surrounding English words and structure by modularly applying the constraints for English. Thus, "borogoves" must be labelled as a noun in order not to violate a noun phrase's exigency for a head noun. In other noun phrases, the extraneous words can be recognised only as adjectives. This is the case for "the manxome foe" and "his vorpal sword", once the following constraints are applied: adjectives must precede nouns, a noun phrase can have only one head noun, determiners are also unique within a noun phrase. In the case of "the slithy toves", where there are two WL words, the constraint that the head noun is obligatory implies that one of these two words is a noun, and the noun must be "toves" rather than "slithy" (which is identified as an adjective as in the two previous examples) in order not to violate the precedence constraint between nouns and adjectives. In other cases we may not be able to unambiguously determine the category, for instance the WL word "frabjous" preceding the English word "day" may remain ambiguous no matter how we parse it, if it satisfies all the constraints either as a determiner or as an adjective[2].

Two of the poem's noun phrases ("the Jubjub bird" and "the Tumtum tree") provide ontological as well as lexical information (under the reasonable assumption that capitalised words must be proper nouns, coupled with the fact that as proper nouns, these words do not violate any constraints). Our adaptation of Womb Grammars includes a starting-point, domain dependent ontology (which could, of course, initially be empty), which can be augmented with such ontological information as the facts that Tumtums are trees and Jubjubs are birds. Similarly, input such as "Vrilligs are vampires" would result in additions to the ontology besides in lexical recognition. It could be that some input allows us even to equate some extraneous words with their English equivalents. For instance, if instead of having in the same poem the noun phrases "his vorpal sword" and "the vorpal blade", we'd encountered "his vorpal sword" and "the cutting blade", we could bet on approximate synonymy between "vorpal" and "cutting" , on the basis of our English ontology having established semantic similarity between "sword" and "blade".

Similarly, extraneous words that repeat might allow a domain-dependent ontology to help determine their meaning. Taking once more the example of "his vorpal sword" and "the vorpal blade", by consulting the ontology besides the constraints, we can not only determine that "vorpal" is an adjective, but also that it probably refers to some quality of cutting objects. It would be most interesting to carefully study under which conditions such ontological inferences would be warranted.

[2]Which precise constraints are defined for a given language subset is left to the grammar designer; those in this paper are meant to exemplify more than to prescribe.

3.3 Dealing with Extraneous Structures

We have said that Womb Grammars figure out the syntax of a target language from that of a source language by "correcting" the latter's syntax to include properties that were violated by the input corpus. Another variant of Womb Grammars, which we call Universal Womb Grammars, does not rely on a specific source language, but uses instead the set of all properties that are possible between any two constituents - a kind of universal syntax. This universal grammar contains contradictory properties, for instance it will state both that a constituent A must precede another constituent B, and that B must precede A. One or both of these properties will be weeded out by processing the input corpus, which is assumed to be correct and representative.

When dealing only with lexical intrusions, our solution discussed in the previous section does not affect the assumption, made by Womb Grammars, that the input corpus is correct: we merely postulate an anonymous category and features, and let constraint solving automatically find out from context which are the "correct" ones (correct in the sense of our multilingual or neologism-creating environment) to associate to an extraneous word.

Extraneous structures, particularly if coexisting with extraneous lexicon, might be more difficult to deal with, because we rely upon the structural constraints being correct in order to infer an unknown category (e.g. the constraint that adjectives must precede nouns helps to determine that the word "vorpal" functions as an adjective in Lewis Carrol's poem). Therefore, in this section we assume there are no extraneous words and we only deal with extraneous structures. We shall then try to combine both approaches.

We assume, with no loss of generality, that the main language is English and that it is being infiltrated with structures of other languages– the same considerations apply if the main language is another one.

One possibility is to use the the Hybrid Womb Grammar approach with the user's mother tongue as target language and English as the source language, thus obtaining a parser for the mixed language, through training a hybrid Womb Grammar with a user-produced representative corpus of sentences. We can then run an input corpus that is representative of the user's talk (e.g. Spanglish) and this will result in a Spanglish grammar adapted to the user in question. Thereafter, this user will be able to create all the neologisms he wants, given that the structures used, although they may be incorrect for either Spanish or English, will be adequately represented in the Spanglish grammar obtained, which is tailored to this user.

3.3.1 Hybrid Parser Generation

3.3.2 The Training Phase

Before being able to parse a user's mixed use of two languages, we propose to obtain a parser for the mixed language, through training a hybrid Womb Grammar with a user-produced representative corpus of sentences. Let L^S (the source language) be the main language used in the text we want to parse, e.g. English. Its syntactic component will be noted L^S_{syntax}, and its lexical component, L^S_{lex}.

Let L^T be the user's mother tongue. We want to obtain the syntax for the user's blending of L^S and L^T. Let us call this mixed language L^M.

Since we have made the assumption that during this training phase we have no extraneous words (that is, no words that do not appear in the lexicon), we have two options: we can either require that the user do not include them in the training phase, so that the target lexicon will be that of English (L^M_{lex}=L^S_{lex}) or we can simply extend the target lexicon to include both the source language's and that of the user's mother tongue (L^M_{lex}=L^S_{lex} union L^T_{lex}). Whichever of these two options we take, let us call the mixed language's lexicon (L^M_{lex}. We can feed a sufficiently representative corpus of sentences in L^T that the user has produced, to a hybrid parser consisting of L^S_{syntax} and L^M_{lex}. This will result in some of the sentences being marked as incorrect by the parser. An analysis of the constraints these "incorrect" sentences violate can subsequently reveal how to transform L^S_{syntax} so it accepts as correct the sentences in the corpus of L_T—i.e., how to transform it into L^T_{syntax}. Figures 1 and 2 respectively show our problem and our proposed solution through Hybrid Parsing in schematic form.

For example, let $L^S = English$ and $L^T = French$, and let us assume that English adjectives always precede the noun they modify, while in French they always post-cede it (an oversimplification, just for illustration purposes). Thus "the blue book" is correct English, whereas in French we would more readily say "le livre bleu".

If we plug the French lexicon and the English syntax constraints into our Womb Grammar parser, and run a representative corpus of (correct) French noun phrases by the resulting hybrid parser, the said precedence property will be declared unsatisfied when hitting phrases such as "le livre bleu". The grammar repairing module can then look at the entire list of unsatisfied constraints, and produce the missing syntactic component of L^T's parser by modifying the constraints in L^S_{syntax} so that none are violated by the corpus sentences.

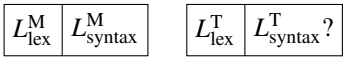

Figure 1: The Problem

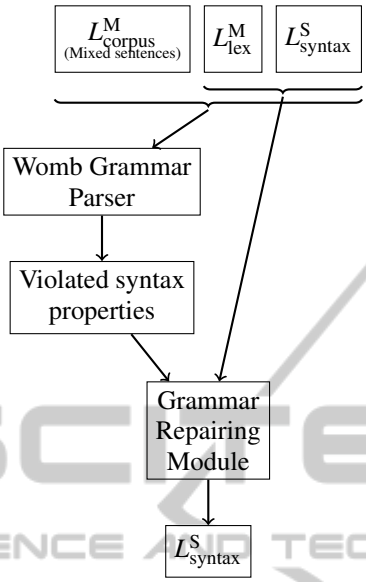

Figure 2: The Solution.

Some of the necessary modifications are easy to identify and to perform, e.g. for accepting "le livre bleu" we only need to delete the (English) precedence requirement of adjective over noun (noted $adj < n$). However, subtler modifications may be in order, perhaps requiring some statistical analysis in a second round of parsing: if in our L^T corpus, which we have assumed representative, *all* adjectives appear after the noun they modify, French is sure to include the reverse precedence property as in English: $n < adj$. So in this case, not only do we need to delete $adj < n$, but we also need to add $n < adj$.

3.4 Inferring Semantic Information

Extracting domain knowledge from text corpora is an active research area which involves several communities (see e.g. (Clark et al., 2012) for an overview). For our purposes we'll focus on the problem building a (partial) hypernym relation graph from textual corpora.

In our context, we are not interested in building a precise structured conceptualisation of a domain but to recognise hypernyms and hyponyms of the extraneous words. Once we are able to recognise the meaning of related words (e.g. using a background source of information like EuroWordNet (Vossen, 2004)) we can classify the missing words and grasp their meaning. For example, searching the web for the exact phrase "a borogove is" returns a snippet containing the sentence "a borogove is a thin shabby-looking bird" which allows us to infer that a "borogove" is a bird.

Different techniques have been developed to optimise the task of acquiring semantic structuring of a domain; however, our problem is much more limited because we are not interested in constructing a complete taxonomy. In particular, the problems of precision and recall will not affect us to the same extent as in the general case.

The fact that we start our search for hypernyms from specific *seed* words and we cannot make strong assumptions on the corpora we are analysing, makes approaches based on hyponym patterns a natural choice (see (Hovy et al., 2009; Snow et al., 2004)). The basic idea is to search the corpora for specific textual patterns which explicitly identify a hyponym relation between terms (e.g., "such authors as $\langle X \rangle$"). Hyponym patterns can be pre-defined or extracted from corpora using known taxonomies (e.g., (Snow et al., 2004)). For our purposes we can reuse known patterns and apply them to the text source being parsed or external sources like Wikipedia or a web search engine (Snow et al., 2006).

4 CONCLUSIONS

We have shown how to use the combined power of Womb grammars plus ontologies in order to make syntactic sense of text for which the grammar we dispose of has only partial information. As well, we have delineated how we could extend these abilities into semantics.

While in this paper we have focused on a specific language's grammar, it might be useful to be able to consult in a second stage the relevant fragment (e.g. that of noun phrases if the extraneous word belongs to one) of a universal grammar. This will be the case for instance if the word that seems not to belong in the text exhibits some property that does not exist in the text's main language. When this is the case, there will be no way to assign for some word a category that is in line with the surrounding ones and results in no more properties being violated.

REFERENCES

Christiansen, H. (2005). CHR grammars. *TPLP*, 5(4-5):467–501.

Christiansen, H. (2011). Adaptable grammars for non-context-free languages. In *Bio-Inspired Models for*

Natural and Formal Languages, pages 33–51. Cambridge Scholars Publishing.

Clark, M., Kim, Y., Kruschwitz, U., Song, D., Albakour, D., Dignum, S., Beresi, U. C., Fasli, M., and De Roeck, A. (2012). Automatically structuring domain knowledge from text: An overview of current research. *Information Processing & Management*, 48(3):552–568.

Dahl, V. and Miralles, J. (2012). Womb grammars: Constraint solving for grammar induction. In Sneyers, J. and Frühwirth, T., editors, *Proceedings of the 9th Workshop on Constraint Handling Rules*, volume Technical Report CW 624, pages 32–40, Department of Computer Science, K.U. Leuven.

Hovy, E., Kozareva, Z., and Riloff, E. (2009). Toward completeness in concept extraction and classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 948–957. Association for Computational Linguistics.

Jackson, Q. T. (2006). *Adapting to Babel: Adaptivity and Context-Sensitivity in Parsing*. Verlag:Ibis Publishing.

Pereira, F. and Warren, D. (1980). Definite clause grammars for language analysis - a survey of the formalism and a comparison with transition networks. *Artificial Intelligence*, 13:231–278.

Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 801–808, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vossen, P. (2004). Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingualindex. *International Journal of Lexicography*, 17(2):161–173.