On Language Acquisition through Womb Grammars

Veronica Dahl, Emilio Miralles, Leonor Becerra-Bonache

On Language Acquisition Through Womb Grammars

Veronica Dahl[1], J. Emilio Miralles[1], and Leonor Becerra[2]

[1] Simon Fraser University, Burnaby, BC, V5A-1S6, Canada,
veronica@cs.sfu.ca, emiralle@sfu.ca
[2] Laboratoire Hubert Curien, Jean Monnet University, 18 rue Benoit Lauras, 42100 Saint-Etienne, France
leonor.becerra@univ-st-etienne.fr

Abstract. We propose to automate the field of language acquisition evaluation through Constraint Solving; in particular through the use of Womb Grammars. Womb Grammar Parsing is a novel constraint based paradigm that was devised mainly to induce grammatical structure from the description of its syntactic constraints in a related language. In this paper we argue that it is also ideal for automating the evaluation of language acquisition, and present as proof of concept a CHRG system for detecting which of fourteen levels of morphological proficiency a child is at, from a representative sample of the child's expressions. Our results also uncover ways in which the linguistic constraints that characterize a grammar need to be tailored to language acquisition applications. We also put forward a proposal for discovering in what order such levels are typically acquired in other languages than English. Our findings have great potential practical value, in that they can help educators tailor the games, stories, songs, etc. that can aid a child (or a second language learner) to progress in timely fashion into the next level of proficiency, and can as well help shed light on the processes by which languages less studied than English are acquired.

Keywords: Womb Grammar Parsing, Language Acquisition, Constraint Order Acquisition, Constraint Based Grammars, Property Grammars, CHRG

1 Introduction

Constraint-based linguistic models, such as HPSG [17] or Property Grammars [1], view linguistic constraints in terms of property satisfaction between categories, rather than in the more traditional terms of properties on hierarchical representations of completely parsed sentences. This view has several advantages, including allowing for mistakes to be detected and pointed out rather than blocking the analysis altogether, and has yielded good results for language analysis and grammar development. Language acquisition is a research area where constraint-based approaches can potentially make important contributions. Applications of constraint-based approaches to processing learner language have

been surveyed in [13], mostly around error detection, as in [3], which represents parsing as a constraint satisfaction problem and uses constraint relaxation with a general-purpose conflict detection algorithm. A recent and original development of constraint-based parsing is the Womb Grammar Parsing paradigm [9], which was designed to induce a target language's constraints on given simple phrases (e.g., noun phrases) from the corresponding constraints of another language called the source, given a corpus of correct phrases in the target language. Womb grammars have proved valuable not only for traditional applications such as analysis, but also for grammar sanctioning and to induce a correct grammar from that of another, related language.

In this article we propose a tailoring of Womb Grammars for another unusual application: that of computer-supported language acquisition, and we exemplify our ideas in terms of a novel application of constraint-based parsing: that of inducing the (incorrect) grammar in use by a person learning a language and detecting the level of proficiency of such a learner.

After presenting our methodological background in the next section, section 3 describes how to use Womb Grammars to detect a child's level of grammatical proficiency and how to induce the linguistic constraints that describe his or her grammar fragment. Section 4 briefly discusses how to adapt our research to discover the learning stages that are followed by languages other than English. Section 5 presents our concluding remarks.

2 Background

Womb Grammar Parsing was designed to induce, from known linguistic constraints that describe phrases in a language called the source, the linguistic constraints that describe phrases in another language, called the target. Grammar induction has met with reasonable success using different views of grammar: a) as a parametrized, generative process explaining the data [16, 14], b) as a probability model, so that learning a grammar amounts to selecting a model from a pre-specified model family [6, 19, 8], and c) as a Bayesian model of machine learning [12].

Most of these approaches have in common the target of inferring *syntactic trees*. As noted, for example, in [2], constraint-based formalisms that make it possible to evaluate each constraint separately are advantageous in comparison with classical, tree-based derivation methods. For instance the Property Grammar framework [1] defines phrase acceptability in terms of the properties or constraints that must be satisfied by groups of categories (e.g. English noun phrases can be described through a few constraints such as precedence (a determiner must precede a noun), uniqueness (there must be only one determiner), exclusion (an adjective phrase must not coexist with a superlative), and so on). Rather than resulting in either a parse tree or failure, such frameworks characterize a sentence through the list of the constraints a phrase satisfies and the list of constraints it violates, so that even incorrect or incomplete phrases will be parsed.

Womb Grammar Parsing follows these general frameworks, but focuses on *generating the constraints* that would sanction the input as correct, rather than on characterizing *sentence acceptability* in terms of (known) linguistic constraints. This is because it was conceived for grammar induction rather than only for parsing sentences, so it can operate on a corpus of sentences deemed correct to generate the set of grammatical constraints (i.e., the grammar description) that *would* result in all constraints being satisfied—i.e., the grammar for the language subset covered by the corpus. Thus, it is ideal for grammar correction and grammar induction, not just for flexible parsing.

More concretely: let L^S (the source language) be a human language that has been studied by linguists and for which we have a reliable parser that accepts correct sentences while pointing out, in the case of incorrect ones, what grammatical constraints are being violated. Its syntactic component will be noted L^S_{syntax}, and its lexical component, L^S_{lex}.

Now imagine we come across a dialect or language called the target language, or L^T, which is close to L^S but has not yet been studied, so that we can only have access to its lexicon (L^T_{lex}) but we know its syntax rules overlap significantly with those of L^S. If we can get hold of a sufficiently representative corpus of sentences in L^T that are known to be correct, we can feed these to a hybrid parser consisting of L^S_{syntax} and L^T_{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}. If we can automate this process, we can greatly aid the work of our world's linguists, the number of which is insufficient to allow the characterization of the myriads of languages and dialects in existence.

An Example. Let $L^S = English$ and $L^T = Spanish$, and let us assume that English adjectives always precede the noun they modify, while in Spanish they always post-cede it (an oversimplification, just for illustration purposes). Thus "a hot burner" is correct English, whereas in Spanish we would more readily say "una hornalla caliente".

If we plug the Spanish lexicon into the English parser, and run a representative corpus of (correct) Spanish noun phrases by the resulting hybrid parser, the said precedence property will be declared unsatisfied when hitting phrases such as "una hornalla caliente". 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.

Some of the necessary modifications are easy to identify and to perform, e.g. for accepting "una hornalla caliente" we only need to delete the (English) precedence requirement of adjective over noun (noted $adj < n$). However, subtler modifications may be in order, requiring some statistical analysis, perhaps 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, Spanish 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 Detecting and Characterizing Grammatical Proficiency

The generative power of Womb Grammars can be used to find out the set of linguistic constraints (i.e. the grammar) in use by a person learning a language. For this end, rather than using the grammar of a known related language as in our example above, we propose that of a Universal Womb Grammar which for each type or phrase lists all *possible* properties or constraints. For instance, for every pair of allowable constituents in a phrase (say noun and adjective in a noun phrase), it would list both possible orderings: $noun < adjective$ and $adjective < noun$. By running a student's input through this universal grammar and deleting any constraints not manifest in the input, we are left with a characterization of the student's proficiency. This grammar represents an individual interlanguage system, which is in line with the general perspective in Second Language Acquisition which brands as a mistake the study of the systematic character of one language by comparing it to another[15].

For instance, the order in which children acquire basic grammatical English constructions is fairly predictable. Table 1 shows a widely accepted morpheme acquisition ordering initially postulated by Brown[4] and corroborated, for instance, by de Villiers and de Villiers[18]. According to these results children acquire a series of 14 morphemes in essentially the *same order*, but *not at the same speed*.

Order	Morpheme
1	Present progressive (-ing)
2-3	Prepositions (in, on)
4	Plural (-s)
5	Irregular past tense
6	Possessive (-'s)
7	Uncontractible copula (is, am are)
8	Articles (the, a)
9	Regular past tense (-ed)
10	Third person present tense, regular (-s)
11	Third person present tense, irregular
12	Uncontractible auxilliary (is, am are)
13	Contractible copula
14	Contractible auxilliary

Table 1. Order of acquisition of morphemes [5]

To use Womb Grammars for this task, we must find an appropriate set of initial constraints (just as in the example in section 2, the constraints of English

are used as the initial set) from which we can weed out those not satisfied at the child's proficiency level. For the present task, we assume a lexicon that includes all features necessary to morphological analysis as per table 1. Morpheme identification is well studied in the literature for uses such as automatic translation, grammar checking, and spelling correction.

Some of the needed initial constraints we are already familiar with. For instance, if we include the constraint that a noun requires a determiner, any input corpus that violates this constraint will prompt its deletion. The resulting output grammar will be incorrect with respect to adult speech, but will adequately characterize the linguistic competence of the child. We can also output the level of proficiency by noting, for instance, that if the output grammar does contain the said requirement constraint, the child is at least at level 8 of the table. In addition, novel uses of the familiar constraints need to come into play. For instance, precedence is typically tested by itself, but when using it to evaluate language proficiency, it needs to be tested *as a condition* to other constraints, e.g. to check that the copula is not contracted when it is the first word in a question (as in "Is he sick?"), we need to check *if* it precedes all other elements in a question, rather than stipulate that it must precede them.

Other necessary constraints need to be tailored to our task. For instance, several of the constraints in the table boil down to whether some particular feature appears in appropriate contexts within the corpus, so all we have to do is check that this is the case. A case in point: the mastery of irregular verbs, as evidenced by their correct inclusion in the corpus, would indicate acquisition at level 5.

Our prototype implementation of incorrect grammar acquisition incorporates both kinds of constraints. This implementation uses CHRG[7], the grammatical counterpart of CHR[11]. The mastery criteria can be set as 90% accuracy, which was the cutoff proposed by Brown[4]. De Villiers and de Villiers[18] introduced a new "ranking method" based on the relative accuracy with which the morphemes were used in obligatory contexts. Many subsequent studies have used some variant of these approaches, and our system can be readily adapted to different methods of rank ordering.

In order for our results to be precise we need that the input corpus be as representative as possible. For instance, if we are targeting 90% success rate as mastery, we need a sufficient number of relevant utterances such that we can reliably translate the number of occurrences into a percentage. Existing child language acquisition corpora, such as the CHILDES database (http://childes.psy.cmu.edu/), could be used for this purpose.

4 Inducing Learning Stages for Languages Other than English

English language acquisition has been very well studied, but there is a dire need for studying the vast majority of languages in existence. An interesting application of Womb Grammars would therefore be to test how much of the English

ordering of morpheme acquisition still holds in any other given language. This could proceed by running our English level proficiency detection system (described in section 3) with the (adequately annotated) morphological lexicon of the language studied. If the results are consistent with those of English, this would indicate a similar learning order. Any differences could be further analyzed in order to find out what the actual acquisition order is in this language. Admittedly this is a rough first approximation, and further work is needed to perfect the approach. There are some constructions that will be learned in wildly different order than in English, for example the plural in Egyptian Arabic is very complex, so children generally do not master it until reaching adolescence.

5 Concluding Remarks

We have argued that Womb Grammar Parsing, whose CHRG implementation is described in [9], is an ideal aid to guide a student through language acquisition by using our proposed Universal Womb Grammar. We have also complimented this prototype with a component that can detect a child's morphological level of proficiency in English.

Our work is most probably also applicable for learning English as a second language, as suggested by studies that show that such learners also progress orderly along the same stages[10]. Unlike previous work, which focuses on machine learning techniques (e.g. [20]), our contribution to quality assessment of utterances in a language being learned proceeds through pointing out which linguistic constraints are being violated. From these, an accurate (while probably incorrect by academic standards) grammar of the users language proficiency can be produced, as well as a set of exercises targeting his or her progress.

Admittedly, much more work is needed for a completely encompassing rendition of this first proposal. For instance, we will need to include phonological and semantic considerations in future work. This process will surely further uncover new constraints that need to be added to the familiar ones for the purposes of our research. Reliable evaluation schemes also need to be devised.

To the best of our knowledge, this is the first time the idea of detecting grammatical performance levels for language acquisition materializes through weeding out constraints from a kind of universal constraint-based grammar fragment. With this initial work we hope to stimulate further research along these lines.

References

1. Blache, P.: Property grammars: A fully constraint-based theory. In: Christiansen, H., Skadhauge, P.R., Villadsen, J. (eds.) CSLP. Lecture Notes in Computer Science, vol. 3438, pp. 1–16. Springer (2004)
2. Blache, P., Guenot, M.L., van Rullen, T.: A corpus-based technique for grammar development. In: Archer, D., Rayson, P., Wilson, A., McEnery, T. (eds.) Proceedings of Corpus Linguistics 2003, University of Lancaster. pp. 123–131 (2003)

3. Boyd, A.A.: Detecting and Diagnosing Grammatical Errors for Beginning Learners of German: From Learner Corpus Annotation to Constraint Satisfaction Problems. Ph.D. thesis, Ohio State University (2012)
4. Brown, R.: A first language: The early stages. George Allen and Unwin (1973)
5. Carroll, D.: Psychology of Language. Thomson/Wadsworth (2008)
6. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: Knight, K., Ng, H.T., Oflazer, K. (eds.) ACL. The Association for Computer Linguistics (2005)
7. Christiansen, H.: Chr grammars. TPLP 5(4-5), 467–501 (2005)
8. Cohen, S.B., Smith, N.A.: Covariance in unsupervised learning of probabilistic grammars. Journal of Machine Learning Research 11, 3017–3051 (2010)
9. Dahl, V., Miralles, J.E.: Womb parsing. In: 9th International Workshop on Constraint Handling Rules (CHR 2012), Budapest, Hungary, September 2012. KU Leuven, Department of Computer Science. pp. 32–40 (2012), Tech. Report CW 624
10. Dulay, H., Burt, M.: Should we teach chidren syntax? In: Language Learning. vol. 23, pp. 245–258 (1973)
11. Frühwirth, T., Raiser, F. (eds.): Constraint Handling Rules: Compilation, Execution, and Analysis (March 2011)
12. Headden, W.P., Johnson, M., McClosky, D.: Improving unsupervised dependency parsing with richer contexts and smoothing. In: HLT-NAACL. pp. 101–109. The Association for Computational Linguistics (2009)
13. Heift, T., Schulze, M.: Errors and Intelligence in Computer-Assisted Language Learning. Parsers and Pedagogues. Routledge, New York, USA (2007)
14. Klein, D., Manning, C.D.: Corpus-based induction of syntactic structure: Models of dependency and constituency. In: Scott, D., Daelemans, W., Walker, M.A. (eds.) ACL. pp. 478–485. ACL (2004)
15. Lakshmanan, U., Selinker, L.: Analysing interlanguage: how do we know what learners know? Second Language Research 14(4), 393–420 (2001)
16. Pereira, F.C.N., Schabes, Y.: Inside-outside reestimation from partially bracketed corpora. In: Thompson, H.S. (ed.) ACL. pp. 128–135. ACL (1992)
17. Pollard, C., Sag, I.A.: Head-driven Phrase Structure Grammars. CSLI, Chicago University Press (1994)
18. de Villiers, J., de Villiers, P.: A cross-sectional study of the acquisition of grammatical morphemes in child speech. In: Journal of Psycholinguistic Research. vol. 2, pp. 267–278 (1973)
19. Wang, M., Smith, N.A., Mitamura, T.: What is the jeopardy model? a quasi-synchronous grammar for qa. In: EMNLP-CoNLL. pp. 22–32. ACL (2007)
20. Yannakoudakis, H., Briscoe, T., Medlock, B.: A new dataset and method for automatically grading esol texts. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 180–189. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002472.2002496