A Parallel Approach to Syllabification

Anca Dinu[1] and Liviu P. Dinu[2]

[1] University of Bucharest, Faculty of Foreign Languages,
5-7 Edgar Quinet, 70106, Bucharest, Romania
anca_radulescu@yahoo.com
[2] University of Bucharest, Faculty of Mathematics and Computer Science,
14 Academiei, 70109, Bucharest, Romania
ldinu@funinf.cs.unibuc.ro

Abstract. In this paper we propose a parallel manner of syllabification introducing some parallel extensions of insertion grammars. We use this grammars in an application to Romanian language syllabification.

1 Introduction

In formal language theory, most of the generative mechanisms investigated are based on the *rewriting* operation. Several other classes of mechanisms, whose main ingredient is the *adjoining* operation, were introduced along the time. The most important of them are *the contextual grammars* (Marcus, 1969), *the tree adjoining grammars* (TAG) (Joshi et al., 1975) and *the insertion grammars* (Galiukschov, 1981), all three of them introduced with linguistic motivations. Contextual grammars were introduced by Marcus (1969) and have their origin in the attempt to bridge the gap between the structuralism and generativism. The insertion grammars (or semi-contextual grammars) are somewhat intermediate between Chomsky context-sensitive grammars (where the non-terminal are rewritten according to specified contexts) and contextual grammars (where contexts are adjoined to specified strings associated with contexts).

In this paper we introduce some parallel extensions of insertion grammars and we use them to propose a parallel manner of word syllabification. Up to now, from our knowledge, most of the formal models of syllabification were treated in a sequential manner (Vennemann (1978), Koskenniemi (1983), Bird and Ellison (1994), Kaplan and Kay (1994), Muller (2002), Dinu (2003)).

This paper is structured as follows: in Section 2 we present *the insertion grammars* and introduce two new variants of them: *parallel insertion grammars* and *maximum parallel insertion grammars*. The syllabification of words, the definition of syllable and an application (Romanian words syllabification) of this approach of syllabification is given in Section 3.

2 Parallel Extensions of Insertion Grammars

For elementary notions of formal language theory, such as *alphabet, concatenation, language, free monoid, lengths of words,* etc. we refer to (Păun, 1997).

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 83–87, 2005.

The basic operation in insertion grammars is the adjoining of strings, as in contextual grammars, not rewriting, as in Chomsky grammars, but the operation is controlled by a context, as in context-sensitive grammars.

Definition 1 (Păun, 1997). *An insertion grammar is a triple $G = (V, A, P)$, where V is an alphabet, A is a finite language over V, and P is a finite set of triples of strings over V.*

The elements in A are called axioms and those in P are called insertion rules.

The meaning of a triple $(u, x, v) \in P$ is: x can be inserted in the context (u, v). Specifically, for $w, z \in V^*$ we write $w \Rightarrow z$ if $w = w_1uvw_2$, $z = w_1uxvw_2$, for $(u, x, v) \in P$ and $w_1, w_2 \in V^*$.

The language generated by G is defined by: $L(G) = \{z \in V^* \mid w \overset{*}{\Rightarrow} z, \text{ for } w \in A\}$.

Here we introduce two parallel extensions of insertion grammars.

Definition 2. *Let $G = (V, A, P)$ be an insertion grammar. We define the parallel derivation denoted $\Rightarrow_p$, by:*

$$w \Rightarrow_p z \text{ iff } w = w_1w_2 \dots w_r, \text{for some } r \geq 2,\ z = w_1x_1w_2x_2w_3 \dots x_{r-1}w_r \text{ and, for all } 1 \leq i \leq r-1, \text{ there is } (u_i, x_i, v_i) \in P \text{ and } \alpha_i, \beta_i \in V^* \text{ such that } w_ix_iw_{i+1} = \alpha_iu_ix_iv_i\beta_i \text{ and } w_i = \alpha_iu_i\ , w_{i+1} = v_i\beta_i.$$

Remark 1. For usual derivation $\Rightarrow$ we use one selector-pair, with no restriction; in parallel derivations the whole string is decomposed into selectors.

Definition 3. *For an insertion grammar $G = (V, A, P)$ we define the parallel derivation with maximum use of insertions (in short, we say maximum parallel derivation), denoted $\Rightarrow_{pM}$, by:*

$$w \Rightarrow_{pM} z \text{ iff } w = w_1w_2 \dots w_s,\ z = w_1x_1w_2x_2w_3 \dots x_{s-1}w_s,\ w \Rightarrow_p z \text{ and there is no } n > s \text{ such that } w = w'_1w'_2 \dots w'_n,\ z' = w'_1x'_1w'_2x'_2w'_3 \dots x'_{n-1}w'_n,\ w \Rightarrow_p z'.$$

Remark 2. The main difference between parallel derivation ($\Rightarrow_p$) and maximum parallel derivation ($\Rightarrow_{pM}$) with respect to an insertion grammar is that in the former we can insert any number of strings in a derivation step and in the later we insert the maximum possible number of strings in a derivation step.

For $\alpha \in \{p, pM\}$, we denote by $L_\alpha(G)$ the language generated by the grammar G in the mode α:

$$L_\alpha(G) = \{z \in V^* \mid w \overset{*}{\Rightarrow}_\alpha z, \text{ for some } w \in A\}.$$

The family of such languages is denoted by INS_α, $\alpha \in \{p, pM\}$.

We give here (without proofs) some results regarding the relations between INS_{pM} and Chomsky hierarchy.

Theorem 1. *INS_{pM} is incomparable to REG and CF, but not disjoint, where REG is the class of regular languages and CF is the class of context free languages.*

Theorem 2. *$INS_{pM} \subset CS$.*

3 On the Syllabification of Romanian Words via Parallel Insertion Grammars

In this section we use the insertion grammars and the maximum parallel insertion derivation to propose a parallel manner of syllabification of words.

Consider an insertion grammar $G = (V, A, P)$ and let $L_{pM}(G)$ be the language generated by G in parallel maximum mode. Set $w \Rightarrow_{pM} z$ a derivation in G, where $w = w_1 w_2 \dots w_s$ and $z = w_1 x_1 w_2 \dots x_{s-1} w_s$.

With respect to the above definitions, we define the syllables of w by:

$$Syl_{pM}(w) = \{w_1, w_2, \dots w_n\}.$$

Consider the Romanian alphabet *RO*={a, ă, â, b, c, d, e, f, g, h, i, î, j, k, l, m, n, o, p, q, r, s, ş, t, ţ, u, v, w, x, y, z} and its partition in vowels and consonants: *RO= Vow* ∪ *Con*, where *Vow*={a, â, ă, e, i, î, o, u, y} and *Con*={b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, ş, t, ţ, v, w, x, z}.

Definition 4. *A word over RO is said to be regular if it contains no consecutive vowels.*

With respect to the above definitions, an insertion grammar for syllabification of Romanian regular words is $G_{syl} = (V_{syl}, A_{syl}, P_{syl})$, whose components are:

1. $V_{syl} = RO \cup \{\$\}$, where "$" is a new symbol that is not in *RO*; "$" is the *syllable boundary marker.*
2. A_{syl} is the set of the regular words over *RO* in Romanian language.
3. $P_{syl} = C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5 \cup C_6 \cup C_7 \cup C_8$ where:
   (a) $C_1 = \{(v_1, \$, cv_2) \mid v_1, v_2 \in Vow, c \in Con\}$
   (b) $C_2 = \{v_1, \$, c_1c_2v_2) \mid v_1, v_2 \in Vow, c_1c_2 \in$ {ch,gh} or $(c_1, c_2) \in$ {b, c, d, f, g, h, p, t} $\times \{l, r\}\}$
   (c) $C_3 = \{(v_1c_1, \$, c_2v_2) \mid v_1, v_2 \in Vow$ and c_1c_2 not as in the precedent case$\}$
   (d) $C_4 = \{v_1c_1, \$, c_2c_3v_2) \mid v_1, v_2 \in Vow,\ c_1c_2c_3 \notin$ {lpt, mpt, mpţ, ncş, nct, ncţ, ndv, rct, rtf, stm}$\}$;
   (e) $C_5 = \{v_1c_1c_2, \$, c_3v_2) \mid v_1, v_2 \in Vow,\ c_1c_2c_3 \in$ {lpt, mpt, mpţ, ncş, nct, ncţ, ndv, rct, rtf, stm}$\}$;
   (f) $C_6 = \{v_1c_1, \$, c_2c_3c_4v_2) \mid v_1, v_2 \in Vow,\ c_1 \in Con,\ c_2c_3c_4 \notin$ {gst, nbl}$\}$;
   (g) $C_7 = \{v_1c_1c_2, \$, c_3c_4v_2) \mid v_1, v_2 \in Vow,\ c_1 \in Con,\ c_2c_3c_4 \in$ {gst, nbl}$\}$;
   (h) $C_8 = \{v_1c_1c_2, \$, c_3c_4c_5v_2) \mid v_1, v_2 \in Vow,\ c_1c_2c_3c_4c_5 \in$ {ptspr, stscr}$\}$;

Example 1. Set the word *lingvistica*. We may have the following parallel derivations:

1. Some parallel derivations:
   $\underbrace{lin}_{w_1}\underbrace{gvisti}_{w_2}\underbrace{ca}_{w_3} \Rightarrow_p lin\$gvisti\$ca$, $\underbrace{lin}_{w_1}\underbrace{gvis}_{w_2}\underbrace{tica}_{w_3} \Rightarrow_p lin\$gvis\$tica$,
   $\underbrace{lingvis}_{w_1}\underbrace{ti}_{w_2}\underbrace{ca}_{w_3} \Rightarrow_p lingvis\$ti\$ca$, $\underbrace{lin}_{w_1}\underbrace{gvis}_{w_2}\underbrace{ti}_{w_3}\underbrace{ca}_{w_4} \Rightarrow_p lin\$gvis\$ti\$ca$,
   etc.
2. The parallel maximum derivation: $\underbrace{lin}_{w_1}\underbrace{gvis}_{w_2}\underbrace{ti}_{w_3}\underbrace{ca}_{w_4} \Rightarrow_{pM} lin\$gvis\$ti\$ca$.

Remark 3. For Romanian words, the only words which can have two different syllabifications are the words ending in "i" (e.g. ochi (noun) and o\$chi (verb)) (Petrovici, 1934). If the final "i" is stressed, the rules $C_1 - C_8$ are applied ,or else the final "i" is considered as a consonant and then the same rules are applied.

Remark 4. In order to syllabicate a *non regular word*, we extracted a set of rules based on the context in which a sequence of 2-5 vowels appears. Thus, we notice that the same group of vowels has an identical syllabification if it has the same letters that precede and/or succeed it (Dinu, 2003). Once we have found a set of rules which characterize the behavior of a sequence of vowels, we use it to extend the grammar G_{syl}.

4 Conclusions

In this paper we have investigated the insertion grammars as generative models for syllabification. We introduced some constraints to the derivation relation, obtaining new classes of insertion languages: insertion languages with parallel derivation (INS_p) and insertion languages with maximum parallel derivation (INS_{pM}). Using the maximum parallel derivation we obtained an efficient method of word syllabification. We analyzed some of the relations between INS_{pM} and the Chomsky hierarchy.

References

1. Bird, S. and T. M. Ellison. One-level phonology: Autosegmental representations and rules as finite automata. *Computational Linguistics* 20, 55-90, 1994
2. Dinu, L.P., An approach to syllables via some extensions of Marcus contextual grammars. *Grammars*, 6 (1), 2003, 1-12.
3. Galiukschov, B.S. Semicontextual grammars (in Russian), *Mat. logica i mat. ling.*, Kalinin Univ. 38-50, 1981
4. *D.O.O.M.*. Ed. Acad., Bucureşti, 1982

5. Joshi, A.K., L.S. Levy, M. Takahashi. Tree adjoining grammars. *J. Computer System Sci.*, 19, 136-163, 1975
6. Kaplan, R.M. and M. Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3), 331-379, 1994
7. Koskenniemi, K. *Two-level morphology: A general computational model for word-form recognition and production.* Doctoral dissertation, University of Helsinki, 1983
8. Marcus, S. Contextual grammars. *Rev Roum. Math. Pures Appl.* 14, 69-74, 1969
9. Müller, K. *Probabilistic Syllable Modeling Using Unsupervised and Supervised Learning Methods* PhD Thesis, Univ. of Stuttgart, Institute of Natural Language Processing, AIMS 2002, vol. 8, no.3, 2002
10. Petrovici, E. Le pseudo *i* final du roumain. *Bull. Linguistique*, 86-97, 1934
11. Păun, Gh. *Marcus Contextual Grammars.* Kluwer, 1997
12. Vennemann, T. Universal syllabic phonology. *Theoretical Linguistics* 5, 2-3, 175-215, 1978