Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar
Elméleti Nyelvészet Doktori Program

The Abstract of the Doctoral Dissertation

A LINGUISTIC AGREEMENT MAPPING-SYSTEM MODEL

(Egy egyeztetésen alapuló nyelvi leképző-rendszer modell)

By

LÁSZLÓ DRIENKÓ

1. Objectives

The primary goal of the dissertation is to investigate how a generalised notion of linguistic agreement can serve as the basis for a linguistic theory. The motivation is that linguistic agreement plays a fundamental role in many languages and such agreement relations cannot be sufficiently managed by substitution-centred theories. We approach language through its underlying agreement relations. A second motivation is that the computational properties of an agreement-based model are quite favourable. In fact, our aim is to achieve linear computational time with respect to the length of the input sequence.

2. The model

The model maps input sequences onto linguistic patterns. The mapping process is guided by checking of the agreement of linguistic features, or attributes, of linguistic elements. The most relevant attributes that we used in the present work include the following:

CAT:	(syntactic) category
PERS:	person
NUM:	number
PHF	phonological form
ARGCASE:	argument case (of a verb), particularly as ARG1CASE: grammatical case of the first argument ARG2CASE: grammatical case of the second argument ARG3CASE: grammatical case of the third argument
PREP	preposition, to check if a verb, noun, adjective, etc. is compatible with a given preposition, particularly as PREP_at PREP_for etc.

GENDER:	gender
DEF:	definiteness
TENSE:	tense
AUX:	auxiliary
AUXN:	auxiliary-noun agreement, to distinguish 'do/does' and 'have/has' from other auxiliary verbs, particularly as AUXN1: for 3rd person singular AUXN2: for non-3rd person-singular
NONF:	non-finite (verb)

Linguistic patterns can be recursive or non-recursive. In the case of recursive patterns agreement relations must be associated with agreement strategies in order that the feature checking process can be unambiguously specified. Strategies can be represented with checking tables. Below we list the strategies that are employed in the present work with examples. For each strategy we give an example input, and a checking table to demonstrate how the strategy could couple elements.

Strategy FIRST-TO-FIRST:

Input: a_1 a_2 a_3 b_1 b_2 b_3

Checking table:

element:	a	b
sequence$_1$	a_1	b_1
sequence$_2$	a_2	b_2
sequence$_3$	a_3	b_3

Strategy LAST-TO-FIRST:

Input: a_1 a_2 a_3 b_1 b_2 b_3

Checking table:

element:	a	b
sequence$_1$	a_1	b_3
sequence$_2$	a_2	b_2
sequence$_3$	a_3	b_1

Strategy ALL-TO-ONE:

Input: a_1 a_2 a_3 b_1

Checking table:

element:	a	b
sequence$_1$	a_1	b_1
sequence$_2$	a_2	b_1
sequence$_3$	a_3	b_1

Strategy FIRST-TO-ONE:

Input: a_1 a_2 a_3 b_1

Checking table:

element:	a	b
sequence$_1$	a_1	b_1

Strategy LAST-TO-ONE:

Input: a_1 a_2 a_3 b_1

Checking table:

element:	a	b
sequence$_1$	a_3	b_1

Strategy ALL-BUT-LAST:

Input: a_1 a_2 a_3 b_1 b_2 b_3

Checking table:

element:	a	b
sequence$_1$	a_1	b_1
sequence$_2$	a_2	b_2

Strategy ORDER:

Input: a_1 a_2 a_3 b_1 b_2 b_3

Checking table:

element:	a b	s(ATTR_ORDER)
sequence$_1$	a_1b_1	s(ATTR_ORDER=x)
sequence$_2$	a_2b_2	s(ATTR_ORDER=x)
sequence$_3$	a_3b_3	s(ATTR_ORDER=x)

Strategy ABORDER:

Input: a_1 a_2 a_3 b_1 b_2 b_3

Checking table:

element:	a b	s(ATTR_ABORDER)
sequence$_1$	a_1 a_2 a_3 b_1 b_2 b_3	s(ATTR_ABORDER =F_{constr}(a_1 a_2 a_3 b_1 b_2 b_3))

Strategy SELF:

Input: a_1 a_2 a_3

Checking table:

element of pattern:	a	a
sequence$_1$	a_1	a_1
sequence$_2$	a_2	a_2
sequence$_3$	a_3	a_3

Strategy OPEN-CLOSE

Input:	a_1	a_2	b_1	a_3	b_2	b_3
	[	[	]	[	]	]

Checking table:

element	a	b
sequence$_1$	a_2	b_1
sequence$_2$	a_3	b_2
sequence$_3$	a_1	b_3

Strategy BRACKETS

Input: a_1 b_1 a_2 b_2 a_3 b_3

Checking table:

element:	a	b	output:
$sequence_1$	a_1	b_1	→ [a_1 b_1] a_2 b_2 a_3 b_3
$sequence_2$	a_2	b_2	→ [a_1 b_1] [a_2 b_2] a_3 b_3
$sequence_3$	a_3	b_3	→ [a_1 b_1] [a_2 b_2] [a_3 b_3]

Strategy MINVAL-TO-ONE

Input: a_1 a_2 a_3 b_1, where a_2 has the least (minimum) value of attribute ATTR

Checking table:

element:	a	b
$sequence_1$	a_2	b_1

Agreement relations are expressed in agreement constraints. A constraint specifies which elements take part in the agreement relation, which attribute value has to be shared by the elements, and which agreement strategy has to be employed. There can be several agreement constraints associated with a pattern. We may prescribe the simultaneous fulfilment of certain constraints, i.e. we may interpret them conjunctively, whereas it is also possible to employ disjunctive patterns representing options as to which agreement conditions should be met. Another dimension of choice is incorporated in conditional agreement constraints, whose application is dependent on certain features of the input.

Patterns consist of representational elements. Representational elements can be atomic and recursive. Only one input element can be mapped onto an atomic representational element, whereas several input elements can be mapped on recursive representational elements. Patterns containing at least one recursive representational element are recursive patterns.

We distinguish two ways of mapping onto patterns: linear and random access. Linear mapping preserves the order of the elements in the input. For random access mapping the order of the input elements is irrelevant. In this work we focus on linear mapping.

The model can resort to a reconstruction mechanism in the event of unknown or underspecified elements in the input. The reconstruction mechanism tries to use the model's 'knowledge' to derive the missing information.

3. Results

We discuss inflection, subcategorization, anaphora, control, raising, unbounded dependencies, Swiss German cross-serial dependency, Bambara reduplication, Old Georgian genitive, Chinese number names, Dutch coordination phrases, and ellipsis.

Inflection, subcategorization, anaphora, control, raising are touched upon mainly to demonstrate the syntactic capacities of the model. Unbounded dependencies, cross-serial dependency, Bambara reduplication Old Georgian genitive, Chinese number names, Dutch coordination phrases, and ellipsis constitute a computational challenge for existing theories.

Swiss German cross-serial dependency (Shieber (1985)), and Bambara reduplication (Culy (1985)), for instance, constitute examples of languages that can only be processed with a machinery more powerful than a context-free parser. Old Georgian genitive (Michaelis and Kracht (1997)), Chinese number names (Radzinski (1991)), Dutch coordination phrases (Groenink (1996)) are known to be beyond the mildly context sensitive class of languages.

Our treatment of the above phenomena within the agreement framework yield linear computational times. Bambara reduplication, and Old Georgian genitive also exemplify how the model can work for other linguistic modules, morphology in those cases.
With the integration of the Reduction Method into the agreement framework we can provide a general treatment of issues like co-ordinated phrases, relative clauses, adjuncts, word order, and discontinuous verbal constituents. The Reduction Method also facilitates the characterisation of human performance limitations numerically, with a single number.
We believe that the syntax of a natural language should not require more than linear time computations. This assumption predicts that languages with more complex processing times should be outside the class of possible natural languages. We show two classes of languages that our model excludes from the class of possible natural languages. One class is characterised by a non-linear agreement checking time. For instance, the data in a)-c) below represent a language where every subject has to agree grammatically with a verb, but not necessarily with its own predicate.

a) I am happy, you are sad, he is hungry
b) I is happy, you am sad , he are hungry
c) *I is happy, you are sad, he is hungry

Another class is exemplified by the data in a)-d) below. Here the sentence can contain any number of adjectives, but only one determiner, one noun, and one verb. The order of the words is irrelevant. Such a language can be brought about by random access mapping onto pattern e).

a) The happy, strong, intelligent dog barks
b) Happy the barks dog strong intelligent
c) *Happy the the dog barks strong
d) *Happy the dog dog barks
e)

Det Adj N V

Besides analysing linguistic phenomena, we also show how various speech deficits can be paralleled by possible breakdowns of certain components of the model. We use the following hypothetical deficits of the computational system to model aphasic defects:

Defect I: defective input elements
Defect II: defective patterns
Defect III: defective attribute value retrieval
Defect IV: defective organisation of the mapping process

In investigating phonological/phonetic errors we use the classification by Blumstein (1995). We analyse syntactic errors in accordance with the data in Grodzinsky (2000) and Bánréti (2001). Reconstruction mechanisms are also assumed throughout our discussion.

4. <u>Problems, future work</u>

The explanatory force of our model lies in the agreement relations as incorporated in agreement patterns. However, as the reader may sometimes rightly feel it, the collections of the various components of the model, – agreement features, strategies, patterns, etc. – do not

seem to constitute a coherent system as yet, even for a single language. The aim of this work was to demonstrate that it is possible to approach language through agreement relations. It requires extensive further research to establish what agreement relations are at work behind natural languages, either individually or cross-linguistically. The more refined selection of the actual agreement constraints, strategies, and patterns – the 'setting of parameters' for the computational system – ought to be determined by future results.
The formal properties of the model can be developed too. For instance, it would be interesting to examine how strategies can be combined in general, or to investigate if there are any principles that could limit the set of possible strategies, besides linear time computability.
It could greatly enhance the 'elegance' of the model if the notion of reduction patterns could somehow be eliminated.
The theoretical advancement of the model would entail improvement of the implementation.
The implementation we describe in the present work reflects the characteristics of an early stage of development.
Our findings with respect to aphasia need further verification, as well. Progress in this area might be slower though, owing to the more limited access to data.

References

Bánréti, Z. (2001) Egyeztetés agrammatikus afáziában: A szintaktikai fa metszése. (Agreement in Agrammatism: Intersecting the Syntactic Tree). Ms.

Blumstein, S. E. (1995). Phonological deficits in aphasia: theoretical perspectives. Hungarian translation: Fonológiai korlátozottságok afáziában: elméleti megközelítések. In Bánréti, Z. (ed.) Nyelvi struktúrák és az agy. Corvina, Budapest, 1999.

Culy, Ch. (1985) The complexity of the vocabulary of Bambara. Linguistics and philosophy *8, pp. 345-351*

Grodzinsky, Y.(2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences* . 23. pp 1-21.

Groenink, A.V. (1996). Mild context-sensitivity and tuple-based generalizations of context-free grammar. Report CS-R9634, Centrum voor Wiskunde en Informatica (CWI), Amsterdam

Michaelis, J, Kracht, M. (1996). *Semilinearity as a Syntactic Invariant*. Paper presented at the conference *Logical Aspects of Computational Linguistics (LACL `96)*, Nancy, September 23-25, 1996. Appeared in C. Retoré (ed.), *Logical Aspects of Computational Linguistics*, pp. 329-345, Springer, Berlin, 1997. Online version: http://tcl.sfs.uni-tuebingen.de/~michael/papers.html

Radzinski, D. (1991) Chinese number-names, tree adjojning languages, and mild context-sensitivity. *Computational Linguistics. 17: 277-299.*

Shieber, S. M. (1985) Evidence against the context-freeness of natural language. *Linguistics and philosophy* 8. 333-343.