From Sentences to Scope Relations and Backward

Gábor Alberti, Márton Károly and Judit Kleiber

Department of Linguistics, University of Pécs, 6 Ifjúság Street, 7624 Pécs, Hungary

Abstract. As we strive for sophisticated machine translation and reliable information extraction, we have launched a subproject pertaining to the revelation of *reference* and *information structure* in (Hungarian) declarative sentences. The crucial part of information extraction is a procedure whose input is a sentence, and whose output is an information structure, which is practically a set of possible operator scope orders (*acceptance*). A similar procedure forms the first half of *machine translation*, too: we need the information structure of the source-language sentence. Then an opposite procedure should come (*generation*), whose input is an information structure, and whose output is an intoned word sequence, that is, a sentence in the target language. We can base the procedure of acceptance (in the above sense) upon that of generation, due to the reversibility of Prolog mechanisms. And as our approach to grammar is "totally lexicalist", the lexical description of verbs is responsible for the order and intonation of words in the generated sentence.

1 Generating and Accepting Hungarian Sentences

As we strive for a sophisticated level of machine translation and reliable information extraction, we have launched a subproject pertaining to the revelation of reference and information structure in declarative sentences.

We are primarily working with data from Hungarian, which is known to be a language with a very rich and explicit information structure (consisting of different types of topics, quantifiers and foci) [1], [2], [3], [4] and an also quite explicit system of four degrees of referentiality [5], [6], [7], [8], including the indefinite specific degree [9] [10]. The kind of input we consider is an ordered set of (Hungarian) words furnished with four stress marks ("unstressed" / "STRESSED" / "**FOCUS-STRESSED**" / "↑CONTRASTIVELY STRESSED↓") – and our program decides if they constitute a well-formed sentence at all, with arguments of appropriate degrees of referentiality and a possible information structure, and delivers these semantic data, including the possible scope orders of topics, quantifiers and foci. We call this direction *acceptance*. We also try to "accept" sequences of words without stress marks: in this case the first step is furnishing them with all possible intonation patterns. A further kind of input is the opposite direction, which can be called *generation*, whose output is an intoned sentence. Generation is based upon the rich lexical description of tensed verbs pertaining to the sentence-internal arrangement and checking of their arguments; and – in harmony with our "totally lexicalist" approach to grammar [11], which can be regarded as a successor of Hudson's [12] *Word Grammar* or

Alberti G., KÃąroly M. and Kleiber J.
From Sentences to Scope Relations and Backward.
DOI: 10.5220/0003017401000111
In *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science (ICEIS 2010)*, page
ISBN: 978-989-8425-13-3

Karttunen's [13] *Radical Lexicalism*, and a formal execution of cognitive ideas similar to those of Croft's [14] *Radical Construction Grammar* – special intra-lexical *generator rules* are responsible for the development of the intricate pre-verbal operator zone of sentences.

In what follows, Sec2 provides a review of the relevant linguistic phenomena, then Sec3 elucidates what we mean by "accepting" potential sentences with or without stress marks and "generating" sentences; and finally we speak about implementation, our theoretical and practical work in progress driven by computational aims.

2 Referentiality Requirements and Information Structure

Hungarian, similar to English in this respect, has an indefinite article (*egy* 'a(n)') and a definite article (*a(z)* 'the') to distinguish different *degrees of referentiality.* This fact seems to suggest two degrees of referentiality, but a closer look to complex facts (in English, in Hungarian and even in Finnish, which lacks articles) proves that there are (at least) three degrees of positive referentiality in the semantic background of Universal Grammar (see example series (1-5) below), besides the lack of referentiality as a fourth degree, which occurs in Hungarian even in the case of countable nouns, as will be shown in (7) below [8]:

Table 1. The four degrees of referentiality (and their expression in Hungarian).

non-referential	referential		
non-specific		specific	
non-definite			definite
∅ (bare singular)	*egy* 'a(n)'	*egy* 'a(n)'	*a(z)* 'the'

The indefinite article is claimed to refer to a specific referentiality: "its referent is a subset of a set of referents already in the domain of discourse" [10] in the English sentence (1e) below, in opposition to the one in the *there* construction (1b):

Example 1. Degrees of referentiality – in English: three (positive) degrees.

a. *There is cock in the kitchen.
b. There is a cock in the kitchen. ⟨+ref, –spec⟩
c. *There is the cock in the kitchen.
d. *Cock is in the kitchen.
e. (?)A cock is in the kitchen. ⟨+spec, –def⟩
f. The cock is in the kitchen. ⟨+def⟩

Even without articles, Finnish can differentiate the three degrees of positive referentiality, by means of word order (⟨-spec⟩: (2a) vs. ⟨+spec⟩: (2b-c)) and number agreement (⟨-def⟩: (2b) vs. ⟨+def⟩: (2c)):

Example 2. Degrees of referentiality – in Finnish: three (positive) degrees, too.

a. Tul - i kaksi suomalais-ta tyttö-ä. ⟨+ref, –spec⟩
come-Past3Sg two Finnish - Part girl-Part 'Two Finnish girls came.'

b. Kaksi suomalais-ta tyttö-ä tul - i. ⟨+spec, –def⟩
two Finnish - Part girl-Part come-Past3Sg
'Two Finnish girls came (out of the four, say, that we expect to come).'

c. Kaksi suomalais-ta tyttö-ä tul - i - **vat.** ⟨+def ⟩
two Finnish - Part girl-Part come-Past-3Pl 'The two Finnish girls came.'

In Hungarian, too, there are similar constructions triggering the Non-Specificity Effect (3b), as well as the Specificity Effect (3e):

Example 3. Degrees of referentiality – in Hungarian: I. Being.

a. *VAN KAkas a KONYhá-ban.
is cock the kitchen - in

b. VAN egy KAkas a KONYhá-ban. ⟨+ref, –spec⟩
is a cock the kitchen - in
'There is a cock in the kitchen.'

c. *VAN a KAkas a KONYhá-ban.
is the cock the kitchen - in

d. *KAkas BENN van a KONYhá-ban.
cock inside is the kitchen - in

e. ?Egy KAkas BENN van a KONYhában. ⟨+spec,–def⟩
a cock inside is the kitchen - in
'A cock is in the kitchen.'

f. A KAkas BENN van a KONYhá-ban. ⟨+def ⟩
the cock inside is the kitchen - in
'The cock is in the kitchen.'

The system is that Patients of verbs expressing being (3a-c), coming into being (4a) and bringing into being (5a) show the Non-Specificity Effect, whilst these verbs regularly have counterparts with Patients showing an opposite, Specificity, effect: see (3e) above, and (4b), (5b) below.

Example 4-5. Degrees of referentiality – in Hungarian: II-III.. Coming / Bringing into being

4a. ÉRkez-ett [*∅/egy/*a] MExikói a KONferenciá-ra.
arrive-Past [∅ / a / the] Mexican the conference - onto
'A Mexican arrived at the conference.'

b. MEG-érkez-ett [*∅ / (?)egy / a] MExikói a KONferenciá-ra.
Perf - arrive - Past [∅ / a / the] Mexican the conference - onto
'A/The Mexican has/had arrived at the conference.'

5a. A GYErek-ek Alakít-ott-ak [*∅ / egy / *az] Énekkar-t a MŰsor-ra.
the child - ren form - Past - Pl [∅ / a / the] choir - Acc the show-onto
'The children formed a choir for the show.'

b. A GYErek-ek MEG-alakít-ott-*ak* [*∅ / (?)egy / az] Énekkar-t a MŰsor-ra.
the child - ren Perf - form - Past-3Pl [∅ / a / the] choir - Acc the show-onto
'The children have formed a/the choir for the show.'

Certain argument positions, thus, undergo positive or negative specificity requirements. Alike, referentiality itself can be studied. As a default, arguments seem to be required to be referential, at least in the post-verbal zone of neutral Hungarian sentences [8]. See the summary in (6):

Example 6. Pos. and neg. ref.-degree requirements in the post-verbal zone of Hung. Sentences.

+ref: (3a), (3d), (4), (5) +spec: (3d-f), (4b), (5b) –spec: (3a-c), (4a), (5a)

Things are even more complicated, however: the above listed requirements can all be neutralized in the pre-verbal operator zone of Hungarian sentences; see (7-9) below. The neutralization of the ⟨+ref⟩ requirement may result in well-formed nominal expressions without any kind of article, as is shown in (7). Such non-referential nominal expressions can occur even in neutral sentences, due to the special pre-verbal position (drawing the stress to itself from the verb stem), occupied by the Patient in (7a), for instance. The neutral meaning content in (8a) below, thus, can be realized in the three word orders listed in the example. Whilst in the case of an adjectival argument, see (8b), the pre-verbal position is the only "shelter" from the Referentiality Requirement, illustrated in (5) above.

Example 7. A few ways of neutralizing positive referentiality-degree requirements in the pre-verbal zone of Hungarian sentences

a. V-modifier
(M) A GYErek-ek *Énekkar-t* alakít-ott-ak a MŰsor-ra.
the child - ren choir - Acc form-Past-3Pl the show-onto
'The children formed a choir for the show.'

b. Focus (F) A GYErek-ek ***É**nekkar-t* alakít-ott - ak a műsor - ra.
the child - ren choir - Acc form-Past-3Pl the show-onto
'It was a choir that the children formed for the show.'

c. Quantifier
(Q) A GYErek-ek *Énekkar-t is* Alakít-ott-ak a MŰsor-ra.
the child - ren choir - Acc also form-Past-3Pl the show-onto
'The children formed also a choir for the show.'

d. Contrastive topic (K) ↑*Énekkar-t*↓ Alakít-hat-tok a MŰsor-ra!
choir - Acc form - can - 2Pl the show-onto
'As for a choir, you are allowed to form it for the show.'

Example 8. Consequence of the neutralization of the Referentiality Requirement.

a. A GYErek-ek *[∅ / egy]* Énekkar-t alakít-ott-ak a MŰsor-ra. ~~⟨+ref⟩~~, ⟨–spec⟩
the child - ren [∅ / a] choir - Acc form - Past-3Pl the show-onto
A GYErek-ek Alakít-ott-ak *egy Énekkar-t* a MŰsor-ra. ⟨+ref⟩, ⟨–spec⟩
the child - ren form - Past - 3Pl a choir - Acc the show-onto
'The children formed a choir for the show.'

b. *A GYErek-ek FESt-ett-ék *ZÖLD-re* a KErítés-t. ⟨+ref⟩, ⟨–ref⟩
the child - ren paint-Past-3Pl green-onto the fence-Acc
A GYErek-ek *ZÖLD-re* fest-ett-ék a KErítés-t. ~~⟨+ref⟩~~, ⟨–ref⟩
the child - ren green-onto paint-Past-3Pl the fence-Acc 'The children painted the fence green.'

Example 9. The neutralization of negative referentiality requirements in the pre-verbal zone of Hungarian sentences (due to another argument's coming into F or K).

a. A **NAGY**szobá-ban van *a kakas*. cf. (3c)
the sitting-room - in is the cock 'It is in the sitting-room that *the cock* is.'

b. **TEG**nap érkez - ett *a mexikói* a konferenciá-ra. cf. (4a)
yesterday arrive-Past3Sg the Mexican the conference-onto
'It was yesterday that *the Mexican* arrived at the conf.'

c. A **GYE**rek-ek alakít-ott-ák *az énekkar-t* a műsor-ra. cf. (5a)
the child - ren form - Past-3Pl the choir-Acc the show-onto
'*The choir* was formed for the show BY THE CHILDREN.'

Table 2 below serves as an illustration of the overall hypothesis concerning the distribution of +/– referentiality restrictions. In a prototypical neutral sentence (type A. below), the sentence-initial topic zone and the postverbal complement zone are devoted to the task of *anchoring* referents, which requires ⟨+ref⟩ arguments, and the tensed verb is the *assertive* center, which contains the new piece of information about the anchored referents. Patients of being, however, straightforwardly belong to the assertive center (see type B). The position immediately preceding the verb stem also belongs to the assertive center, and hence provides "shelter" for genuinely non-referential arguments (type C). What is common in types D and E, is that some assertive operator appears in the sentence (e.g. a focus), which draws the assertive center to itself from other zones of the sentence. As a consequence, positive referentiality-degree requirements are neutralized in the new assertive zone (D), whilst negative ones are neutralized elsewhere in the sentence structure (E).

Table 2. Anchoring (grey) and assertive pieces of information in different sentence types.

A. Prototypical neutral sentence with *anchoring* arguments and an *assertive* verb (5b):			
A GYErek-ek$_{+ref}$ the child-ren	MEG-alakít-ott-ák Perf-form-Past-Pl3	az Énekkar-t$_{+ref}$ a MŰsor-ra$_{+ref.}$ the choir-Acc the show-onto	
B. Neutral sentence with an argument expressing being, hence belonging to the *assertive* zone (5a):			
A GYErek-ek$_{+ref}$ the child-ren	Alakít-ottak egy Énekkar-t$_{+ref\ -spec}$ form-Past-3Pl a choir - Acc		a MŰsor-ra$_{+ref.}$ the show-onto
C. Neutral sentence with an argument expressing being in the preverbal modifier position, hence belonging to the *assertive* zone (8a):			
A GYErek-ek$_{+ref}$ the child-ren	∅/egy Énekkart$_{-spec}$ alakít-ott-ak ∅/a choir-Acc form-Past-3Pl		a MŰsor-ra$_{+ref.}$ the show-onto
D. Focused sentence I: the assertive zone is occupied by a Non-Specificity Effect argument, due to its focus status (while the verb gets out of the assertive zone also due to the focus construction) (7b):			
A GYErek-ek$_{+ref}$ the child-ren	Énekkar-t$_{+ref}$ choir-Acc	alakít-ott-ak a műsor-ra. form-Past-3Pl the show-onto	
E. Focused s. II: the assertive zone is occupied by a focused constituent, while the verb and a Specificity-Effect argument get out of the assertive zone (due to the focus construction) (9d):			
A **GYE**rek-ek the child-ren	alakít-ott-ák az énekkar-t$_{-spec}$ a műsor-ra. form-Past-3Pl the choir-Acc the show-onto		

3 Generating Sentences, Accepting (Intoned) Word Sequences

The crucial part of *information extraction* is a procedure whose input is a sentence, and whose output is an information structure, which is practically a set of possible

operator scope orders (*acceptance*). A similar procedure forms the first half of *machine translation*, too: we need the information structure of the source-language sentence. Then an opposite procedure should come (*generation*), whose input is an information structure, and whose output is an intoned word sequence, that is, a sentence in the target language. Now let us consider this latter procedure, because the former procedure will be based upon it.

As our approach to grammar is "totally lexicalist" [11], similar to our earlier attempts [15], the lexical description of a verb is responsible for the order and intonation of words in the *generated* sentence. What is demonstrated in (10a) below is the requirement, registered in the *core lexicon*, that the subject of *alakít* 'form' should be the (stressed) topic in the initial part of the sentence ('⟨1,T⟩'), the object should occupy the (stressed) verbal modifier position ('⟨2,M⟩') (with an unstressed verb stem following it), and the *-rA* expression should remain in an (also stressed) post-verbal argument position. The numbers provide a scope order, which is still, in a neutral sentence, practically irrelevant. Let us call this lexical rule the *generator*. In (10b), the default generator in the core lexicon requires the *-rA* argument to occupy the verbal modifier position. In (10c) the Hungarian neutral word order requires a generator that ensures that the prefix will occupy the position next to the verb stem and the (non-agentive) subject will remain in a post-verbal A position.

Example 10. A few Hungarian lexical items with default scope order in the core lexicon.

a. FORM($Arg_{\varnothing}$, Arg_{-t}, Arg_{-ra}) (*default generator* in the core lexicon)
⟨⟨1,T⟩,⟨2,M⟩,⟨3,A⟩⟩
A GYErekek Énekkart alakítottak a MŰsorra. (7a)
'The children have/had formed a choir for the show.'

b. PAINT($Arg_{\varnothing}$, Arg_{-t}, Arg_{-ra})
⟨⟨1,T⟩,⟨3,A⟩,⟨2,M⟩⟩
A GYErekek ZÖLDre festették a KErítést. 'The children painted the fence green.' (8b)

c. ARRIVE(Prefix, $Arg_{\varnothing}$, Arg_{-ra})
⟨⟨1,M⟩,⟨2,A⟩,⟨3,A⟩⟩
MEGérkezett a MExikói a KONferenciára. 'The Mexican arrived at the conference.' (4b)

Two further types of generators produce verb variants, located in an extended lexicon. What is shown in the first row of (11a) is that an *extending* generator inserts certain adjuncts in the argument structure as "fake arguments". Then an *inducing* generator is shown in the second row, responsible for transporting the two fake arguments into the sentence-initial topic zone, the subject into a quantifier position, and the object into the focus. This generator is responsible for the distribution of appropriate stresses. (11b-c) also show an extending generator and two inducing ones. Formula '⟨1,K⟩' refers to a contrastive topic position in the initial part of the pre-verbal Hungarian operator zone.

Example 11. Lexical rules extending argument structures and inducing non-neutral ones.

a. FORM($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$)$\wedge\langle Arg_{Time}, Arg_{Place}\rangle$ (*extending generator*)
$\langle\langle 3,Q\rangle,\langle 4,F\rangle,\langle 5,A\rangle\rangle,\langle\langle 1,T\rangle,\langle 2,T\rangle\rangle$ (*inducing generator*)
TEGnap a KLUB-ban a GYErek-ek is **É**nekkar-t alakít-ott-ak a műsor-ra.
yesterday the club - in the child - ren also choir - Acc form - Past-3Pl the show-onto
'It was a choir that yesterday in the club the children, too, formed for the show.'

b. FORM($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$)$\wedge\langle Arg_{Place}\rangle$
$\langle\langle 2,F\rangle,\langle 3,M\rangle,\langle 4,A\rangle\rangle,\langle\langle 1,K\rangle\rangle$
A ↑KLUB-ban↓ a **GYE**rek-ek alakít-ott-ak énekkar-t a műsor-ra.
the club - in the child - ren form-Past-3Pl choir - Acc the show-onto
'As for the club, a choir was formed there for the show BY THE CHILDREN.'

c. PAINT($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$)
$\langle\langle 1,T\rangle,\langle 2,Q\rangle,\langle 3,F\rangle\rangle$ 'The children painted each fence GREEN.'
A GYErek-ek MINdegyik KErítés-t **ZÖLD**-re fest-ett-ék.
the child - ren every fence - Acc green-onto paint-Past-3Pl

What a generator produces is generally a *set* of sentences, typically ones with different word orders, arranged in a preference order. In Hungarian, it is the *quantifier* that is responsible for this phenomenon, because a quantifier can choose between occupying its preverbal operator position according to the scope order (σ_1, σ_{10}, σ_{11}, σ_{20}, σ_{30}) or remaining in the post-verbal zone (σ_2, σ_3, σ_{20}, σ_{30}, σ_{40}, σ_{50}).

Example 12. Generating intoned sentences: $\nu \rightarrow \langle\sigma_1, \sigma_2 \ldots \sigma_k\rangle$.

a. PAINT($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$)
ν: $\langle\langle 3,Q\rangle,\langle 1,F\rangle,\langle 2,A\rangle\rangle \rightarrow \langle\sigma_1, \sigma_2, \sigma_3\rangle$
σ_1: MINdegyik KErítés-t ZÖLD-re fest-ett-ék a GYErek-ek.
each fence-Acc green-onto paint-Past-3Pl the child-ren
σ_2: ZÖLDre festették a GYErekek MINdegyik KErítést.
σ_3: ZÖLDre festették MINdegyik KErítést a GYErekek.
'Each fence has been painted green by the children.'

b. PAINT($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$)
ν: $\langle\langle 1,Q\rangle,\langle 2,Q\rangle,\langle 3,M\rangle\rangle \rightarrow \langle\sigma_{10}, \sigma_{20}, \sigma_{30}, \sigma_{40}, \sigma_{50}\rangle$
σ_{10}: A GYErekek is ('also') MINdegyik KErítést ZÖLDre festették.
σ_{20}: A GYErekek is ZÖLDre festették MINdegyik KErítést.
σ_{30}: MINdegyik KErítést ZÖLDre festették a GYErekek is.
σ_{40}: ZÖLDre festették a GYErekek is MINdegyik KErítést.
σ_{50}: ZÖLDre festették MINdegyik KErítést a GYErekek is.
'The children, too, have painted each fence green.'

c. PAINT($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$) 'Each fence has been painted green also by the children.'
ν: $\langle\langle 2,Q\rangle,\langle 1,Q\rangle,\langle 3,M\rangle\rangle \rightarrow \langle\sigma_{11}, \sigma_{30}, \sigma_{20}, \sigma_{50}, \sigma_{40}\rangle$
σ_{11}: MINdegyik KErítést a GYErekek is ZÖLDre festették.

The generated set can also be empty ($\lambda = \langle\rangle$). The problem in (13a) below is that an adjectival (hence, $\langle$–ref$\rangle$) argument cannot accept the argument position suggested by the inducing generator (5b). In (13b) the inducing generator doubly violates the Referentiality Requirement (5), which is not neutralized in a (non-contrastive) topic position (7). (13c) illustrates the violation of the possible operator order (in Hungarian), which is as follows (7): {T, K}* $\wedge$ {Q, F}* $\wedge$ (M)A*.

Example 13. Generating an empty set of intoned sentences: $\nu \rightarrow \langle \sigma_1, \sigma_2 \ldots \sigma_k \rangle = \lambda$.

PAINT($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$) the child - ren paint-Past-3Pl green-onto the fence-Acc

a. ν_1: $\langle\langle 1,T\rangle,\langle 3,A\rangle,\mathbf{\langle 2,A\rangle}\rangle \rightarrow \lambda$: *A GYErek-ek FESt-ett-ék *ZÖLD-re* a KErítés-t.

b. ν_2: $\langle\mathbf{\langle 1,T\rangle},\mathbf{\langle 3,A\rangle},\langle 2,M\rangle\rangle \rightarrow \lambda$: *GYErek-ek ZÖLD-re fest-ett-ek KErítés-t.
child - ren green-onto paint-Past-3Pl fence-Acc

c. ν_3: $\langle\mathbf{\langle 1,Q\rangle},\mathbf{\langle 2,T\rangle},\langle 3,M\rangle\rangle \rightarrow \lambda$: *A GYErek-ek is a KErítés-t ZÖLD - re fest-ett-ék.
the child - ren also the fence-Acc green-onto paint-Past-3Pl

Acceptance of an intoned word sequence is the reverse of generation, and can be based upon the latter: we should collect the potential core-lexical verbs, and then collect potential generators, and finally test all combinations in some effective way: see (14) below. If the input contains no reference to intonation, the first step of acceptance is to furnish the word sequence with all potential stress patterns – in some effective way, of course. (15) will serve with some illustration.

Example 14. Accepting scope orders (input: intoned sentence): $\sigma \rightarrow \langle \nu_1, \nu_2 \ldots \nu_k \rangle$.

a. σ_{10}: A GYErekek is MINdegyik KErítést ZÖLDre festették. $\sigma_{10} \rightarrow \langle \nu_1 \rangle$; see (12) above

PAINT($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$)

ν_1: $\langle\langle 1,Q\rangle,\langle 2,Q\rangle,\langle 3,M\rangle\rangle$

b. σ_{11}: MINdegyik KErítést a GYErekek is ZÖLDre festették. $\sigma_{11} \rightarrow \langle \nu_2 \rangle$

ν_2: $\langle\langle 2,Q\rangle,\langle 1,Q\rangle,\langle 3,M\rangle\rangle$

c. σ_{20}: A GYErekek is ZÖLDre festették MINdegyik KErítést. $\sigma_{20} \rightarrow \langle \nu_1, \nu_2 \rangle$

d. σ_{30}: MINdegyik KErítést ZÖLDre festették a GYErekek is. $\sigma_{30} \rightarrow \langle \nu_2, \nu_1 \rangle$

e. σ': *GYErekek **ZÖLD**re MINdegyik KErítést festették. $\sigma' \rightarrow \varnothing$

Example 15. Accepting scope orders (input: a word order with no intonation): $\kappa \rightarrow \langle \sigma_1, \sigma_2 \ldots \sigma_n \rangle$, where $\sigma_1 \rightarrow \langle \nu_{1,1}, \nu_{1,2} \ldots \nu_{1,k_1} \rangle \ldots \sigma_n \rightarrow \langle \nu_{n,1}, \nu_{n,2} \ldots \nu_{n,k_n} \rangle$.

κ: a gyerek-ek is zöld-re fest-ett-ék mindegyik kerítés-t
the child - ren also green-onto paint-Past-3Pl each fence-Acc

$\kappa \rightarrow \langle \sigma_{20}, \sigma_{22}, \sigma_{23}, \ldots \rangle$; PAINT($Arg_{\varnothing}$, $Arg_{\text{-t}}$, $Arg_{\text{-ra}}$)

a. σ_{20}: A GYErekek is ZÖLDre festették MINdegyik KErítést.
'The children, too, painted each fence green.'

$\sigma_{20} \rightarrow \langle \nu_1, \nu_2 \rangle$; ν_1: $\langle\langle Q,1\rangle, \langle Q,2\rangle, \langle M,3\rangle\rangle$; ν_2: $\langle\langle Q,2\rangle, \langle Q,1\rangle, \langle M,3\rangle\rangle$

b. σ_{22}: A GYErekek is **ZÖLD**re festették MINdegyik KErítést.
'It is green that the children, too, painted each fence.'

$\sigma_{22} \rightarrow \langle \nu_{12}, \nu_{22} \rangle$; ν_{12}: $\langle\langle Q,1\rangle, \langle Q,2\rangle, \langle F,3\rangle\rangle$; ν_{22}: $\langle\langle Q,2\rangle, \langle Q,1\rangle, \langle F,3\rangle\rangle$;
$^{?}\nu_{23}$: $\langle\langle Q,1\rangle, \langle Q,3\rangle, \langle F,2\rangle\rangle$

c. σ_{23}: $^{??}$↑A GYErekek is↓ ZÖLDre festették MINdegyik KErítést.
'What is true for a set containing children and other people, is nothing else but that it is green that they painted each fence.'

$\sigma_{23} \rightarrow \langle \nu_{13}, \nu_{23} \rangle$; $^{?}\nu_{13}$: $\langle\langle K,1\rangle, \langle Q,2\rangle, \langle F,3\rangle\rangle$; $^{??}\nu_{23}$: $\langle\langle K,1\rangle, \langle Q,3\rangle, \langle F,2\rangle\rangle$

d. σ_{44}: *A **GYE**rekek is ZÖLDre festették MINdegyik KErítést. $\sigma_{44} \rightarrow \varnothing$

The last example in this section concerns translation between Hungarian and English. As English word order is very strict, what corresponds to an inducing generator in Hungarian is not simply the same generator but its combination with what we also regard as a lexical generator: one producing new argument structure versions by passivization or dative shift. This approach is also supported by Croft's ([14], Section 8) typological analyses: the best way of understanding the numerous

intermediate forms across languages of the world between the active version and the English-type standard passive form ('voice continuum') is in terms of demands for topicalization pertaining to different argument positions.

Example 16. Hungarian operators ~ English argument structure versions.

a. GIVE($Arg_{\varnothing}$, Arg_{-nAk}, Arg_{-t}) — GIVE($Arg_{\varnothing}$, Arg_{Obj_1}, Arg_{Obj_2})
ν_2: ⟨⟨1,T⟩,⟨3,A⟩,⟨2,F⟩⟩ — ν_2: ⟨⟨1,T⟩,⟨3,A⟩,⟨2,F⟩⟩
PÉter egy **KÖNY**v-et ad-ott Mari-nak. — Peter gave Mary a **BOOK**.
Peter a book - Acc give-Past3sg Mary-Dat

b. GIVE($Arg_{\varnothing}$, Arg_{-nAk}, Arg_{-t}) — GIVE($Arg_{\varnothing}$, Arg_{Obj_1}, Arg_{Obj_2})
ν_3: ⟨⟨1,T⟩,⟨2,F⟩,⟨3,A⟩⟩ — ⟨$Arg_{\varnothing}$, Arg_{to}, Arg_{Obj}⟩ + ν_3
PÉter **MA**ri-nak ad-ott egy könyv-et. — Peter gave a book to **MA**ry.
Peter Mary-Dat give-Past3sg a book - Acc

c. GIVE($Arg_{\varnothing}$, Arg_{-nAk}, Arg_{-t}) — GIVE($Arg_{\varnothing}$, Arg_{Obj_1}, Arg_{Obj_2})
ν_4: ⟨⟨2,F⟩,⟨1,T⟩,⟨3,A⟩⟩ — ⟨Arg_{by}, $Arg_{\varnothing}$, Arg_{Obj}⟩ + ν_4
MAri-nak **PÉ**ter ad-ott egy könyv-et. — Mary was given a book by **PE**ter.
Mary-Dat Peter give-Past3sg a book-Acc

d. GIVE($Arg_{\varnothing}$, Arg_{-nAk}, Arg_{-t}) — GIVE($Arg_{\varnothing}$, Arg_{Obj_1}, Arg_{Obj_2})
ν_5: ⟨⟨2,F⟩,⟨3,A⟩,⟨1,T⟩⟩ — ⟨Arg_{by}, Arg_{to}, $Arg_{\varnothing}$⟩ + ν_5
A KÖNYv-et **PÉ**ter ad - t - a Mari-nak. — The book was given to Mary by **PE**ter.
The book - Acc Peter give-Past-3sg Mary-Dat

We note at this point, following our anonymous reviewer's advice: what (16) illustrates is a simplified map of the relevant phenomena; as 'intonation' includes not only stress/prominence but also, as a separate set of choices, tone (e.g. falling/rising). We say that an 'intoned' word sequence is generated, but it would be more accurate to say that it is a 'word sequence with prominence'. An exception is the Hungarian contrastive topic with its rising-falling tone, which we consider; but the special tonal pattern of yes/no questions and ironic performances has not been considered yet.

In (spoken) English prominence alone can signal information structure, in a way that is not possible in Hungarian, e.g. (16c) could be expressed in English as ***PE**ter gave Mary a book*. We assume in our simplified approach that tonic focus is always on the last lexical word in the sentence, being aware of the fact that it is only the unmarked case. Considering the marked cases is a task postponed to future research.

4 Implementation

There are only a few works pertaining to deep parsing of scope order and referentiality in connection with word order and intonation. An (excellent) example is shown by Traat and Bos [16], but we are the first to attempt to build a similar system for Hungarian (whose relevant and advantageous properties are discussed in Sec. 2).

First of all we analyze sentences phonologically and morphologically. Our approach is "totally lexicalist" – grammars based on "total lexicalism" need *not* build phrase structures [11], [12], [13], [14]. In our system, word order is handled by *rank parameters*, instead. In general, we use whole numbers from 1 to 7 for ranks, 1 being

the strongest; by default, it means direct adjacency. Our lexicon contains *morphemes* instead of words and they can search for any other morpheme inside or outside the word which they are part of. In our approach, the only difference between morphology and syntax is that in the syntactic subsystem, the program searches for morphemes being in a certain grammatical relation in two *different* words. But because of this, morphological and syntactic rank parameters are handled separately.

The lexicon is extendable in multiple ways. Apart from adding more words and morphemes, new features can be added to the data structure, thus improving the precision of the analysis. Third, the *core lexicon* can be extended via generating new lexical elements by rules of generation, thus forming the *extended lexicon*. The core lexicon contains all the basic properties of a morpheme including its default behavior (e.g. argument structure of a verb). This applies primarily for intonation: the core lexicon contains the property of "stress" (some words cannot be stressed at all while others are stressed in a neutral sentence) – but it can be overwritten if the sentence is not neutral. For instance, the entity with the property value "focus-stressed" is generated automatically and inserted into the extended lexicon.

Obviously, it is not very effective to try out all of the possible intonation schemes on a long sentence. However, there are two things which must be taken into account: apart from morphemes (such as articles) which can never be stressed or focus-stressed, intonation of arguments before and after the verb is constrained.

By default, the normal Hungarian argument position is *after* the verb. There are, however, arguments which prefer being in the verbal modifier or the topic position. These preferences should be stored in the core lexicon along with the arguments, as was illustrated in (10) in Section 3. If the sentence does not fit into the preferred schema, it is best to use heuristics to create the appropriate generator (11-12).

Prolog (a language for logical programming), including Visual Prolog 7 (which is probably the most elaborate version of Prolog and which we use), usually allows writing predicates which can be invoked "backwards" (*generation* (12) vs. *acceptance* (14)). We suppose that the evaluation of predicates can be reversed. Based on this principle, the future machine translator can be symmetrical. By using the keyword `anyflow`, all arguments of the predicate can be used for both input and output, allowing even entire programs to be executed "backwards". The keywords `procedure`, `determ`, `multi`, `nondeterm` etc. describe whether the predicate can fail (a procedure must always succeed), have multiple backtrack points (`nondeterm`) or not (`determ`).

Let us turn to the particular phases of parsing (*acceptance*).

Phase 0. Before taking intonation etc. into account, all words are segmented and analyzed phonologically and morphologically, allowing the class of each word to be determined. Practically, the last class-changing morpheme (derivative affix) has the relevant "class" output feature. The input and output classes are stored in the core lexicon of every class-changing morpheme.

Phase 1. During the actual syntactic analysis, arguments of a verb are searched with rank 7 (weak). This is a bi-directional search: all non-predicative nominal expressions look for their predicate, too, and if appropriate, the result is stored in the memory of the computer. The adverbial adjuncts search for the verb with rank 7, too, but if it is found, the extending generator must be invoked to modify the default argument structure of the verb and insert the form into the extended lexicon (11a-b).

Phase 2. The inducing generator tries to determine the discourse function – T, Q, F, M, A, see (7) – of the arguments of a verb by creating all possible patterns and trying to apply them onto the sentence. If intonation is present in the input, it can be considered here, making the analysis faster. But shortly, if the Hungarian operator order is violated, the analysis is apparently wrong and the backtracking mechanism of Prolog should be activated before we reach the end of the sentence (13c).

Phase 3. As for the referentiality degrees (6), two of them (non-referential and definite) can be handled during Phase 1. First of all we suppose that all nouns are non-referential by default. Since the article searches the noun, it changes the features 'positive ref. degree' and 'negative ref. degree' of the noun and the new form should be inserted into the extended lexicon.

The only remaining problem is (non-)specificity. Let us take (3). If a non-definite determinant is present, this can be determined only (partly) during or after Phase 2. If no other constraints exist and a verbal modifier is present in a neutral sentence, the argument is specific. So if we generate two instances of *egy mexikói* 'a Mexican', one with features 'spec' and 'non-def' and one with 'non-spec' and 'ref', the analysis of (3b) must fail with the latter. Of course, the generated instances must be inserted into the extended lexicon. If the verb has a modifier, its argument structure *differs* from that of the *same* verb without a modifier because of the specificity criteria. An easier but slower way is to insert two *egy*'s 'a(n)' into the core lexicon, one being specific and one not. If the core lexicon has two morphemes with the same body, both will be taken into account from the beginning (see Phase 0) and this may slow down the analysis because if the result of Phase 1 or 2 is wrong, much of the analysis, including parts of Phase 0 will start over because even the most basic unifications (belonging to the same unit in the *core lexicon*) will have to be broken up.

5 Conclusions

Very few computational systems exist which aim at deep parsing of scope order and referentiality in connection with word order and intonation, although this is a crucial step in reliable information extraction and sophisticated machine translation. We strive for building a system based – at first – primarily upon Hungarian, which is famous for expressing scope relations (in its pre-verbal operator zone) in an explicit way [2], [3] and referentiality degrees also in a quite straightforward manner [5], [8]. We also began to extend our approach to other languages (e.g. English), pointing to similarities, differences and correspondences, for instance, between Hungarian word-order changing operations and English argument-structure changing operations.

To achieve this goal in our totally lexicalist approach [11], the implementation requires a double-layered lexicon. This consists of a core lexicon containing the default behaviour of morphemes (e.g. verb stems) and an extended lexicon, whose elements are generated by a generic lexical-rule module. The elements of the core lexicon are responsible for accounting for neutral sentences, whilst those of the extended lexicon are to handle sentences with different topic, quantifier and focus constructions, where referentiality requirements are often different from those in neutral sentences. We thus primarily generate neutral and non-neutral sentences from

the elements of the double-layered lexicon, and, based upon this generation in our Prolog environment, we extract the information structure of +/– intoned sentences.

Acknowledgements

We are grateful to OTKA60595 for their contribution to all our costs at NLPCS 2010.

References

1. Kiefer, F., É. Kiss, K. (eds.): The Syntactic Structure of Hungarian. Syntax and Semantics, Vol. 27. Academic Press, New York (1994)
2. É. Kiss, K.: Hungarian Syntax. Cambridge University Press, Cambridge (2001)
3. Szabolcsi, A.: Strategies for Scope Taking. In: Szabolcsi, A. (ed.): Ways of Scope Taking, SLAP 65. Kluwer, Dordrecht (1997) 109–154
4. Alberti, G., Medve, A.: Focus Constructions and the "Scope–Inversion Puzzle" in Hungarian. In: Approaches to Hungarian, Vol.7. Szeged (2000) 93–118
5. Szabolcsi, A.: From the Definitness Effect to Lexical Integrity. In: Abraham, W., de Meij, S. (eds.): Topic, Focus, and Configurationality. John Benjamins, Amsterdam (1986) 321–348
6. É. Kiss, K.: Definiteness Effect Revisited. In: Appr. to Hung., Vol. 5. (1995) 63–88
7. Kálmán, L.: Definiteness Effect Verbs in Hungarian. In: Appr. to Hung., Vol. 5. (1995) 221–242
8. Alberti, G.: Restrictions on the Degree of Referentiality of Arguments in Hungarian Sentences. In: Acta Linguistica Hungarica, Vol. 44/3-4. (1997) 341–362
9. de Jong, F., Verkuyl, H.: Generalized Quantifiers: the Properness of Their Strength. In: van Benthem, J., ter Meulen, A. (eds.): GRASS 4. Foris, Dordrecht (1984)
10. Enç, M.: The semantics of specificity. In: Linguistic Inquiry, 22. (1991) 1–25
11. Alberti, G., Kleiber, J. Viszket, A.: GeLexi project: Sentence Parsing Based on a GEnerative LEXIcon. In: Acta Cybernetica, Vol 16. (2004) 587–600
12. Hudson, R.: Word Grammar. Blackwell, Oxford (1984)
13. Karttunen, L.: Radical Lexicalism. Report No. CSLI 86-68. Stanford (1986)
14. Croft, W.: Radical Construction Grammar. Syntactic Theory in Typological Perspective. Oxford University Press, Oxford (2001)
15. Alberti, G. Kleiber, J.: The GeLexi MT Project. In: Hutchins, J. (ed.): Proceedings of EAMT 2004 Workshop (Malta), Univ. of Malta, Valletta (2004) 1–10
16. Traat, M. Bos, J.: Unificational Combinatory Categorial Grammar. In: Proceedings of the 20th International Conference on Computational Linguistics. Geneva (2004)