On the Hungarian examples database

Gergely Pethő 2006

Research supported by ESRC award RES-000-22-0941

This section explains the database of examples that served as the primary source of data used in our research. The following aspects of the database are explained in detail: Part I briefly introduces the Hungarian National Corpus, which was used as the source for the examples to be included in the database. Part II explains the searches that were carried out in the corpus to gather the needed data. Part III discusses the manual processing of the automatically retrieved data, in particular the classification of the retrieved examples according to largely theory-neutral descriptive categories. The descriptive categories used are explained in detail. Part IV examines the information types included in the database entries. Finally, part V contains a short summary of the size and content of the database.

I. ON THE HUNGARIAN NATIONAL CORPUS

The empirical basis of our research was a collection of corpus examples, compiled from the Hungarian National Corpus (Magyar Nemzeti Szövegtár, HNC). The Hungarian National Corpus is a very large collection of electronic texts (consisting of approximately 160 million words) which is partially balanced, following the model of the British National Corpus: It mainly contains written texts, but some transcripts of spoken language are also represented. The written material contains texts from many different genres, including Hungarian literature, scientific and technical texts, archives of an internet forum, and, most importantly, newspaper articles. Like in the case of the BNC, the aim of compiling a balanced corpus of Hungarian was to provide an optimally representative sample of present-day use of the given language, both its spoken and written varieties. The corpus is fully part-of-speech-tagged and includes morphological annotation (for nominal case forms, verbal inflection etc.).

Material from the HNC was retrieved using the web search interface of the corpus, which allows users to search for part-of-speech and morphological tags in addition to particular words and word forms. Therefore, simple searches can be carried out for certain syntactic structures, although search results contain a large percentage of irrelevant examples, partly because of mistakes of the (automatic) annotation of the texts, and partly because only linear sequences of elements can be specified as search terms, but no hierarchically structured expressions.

II. SEARCH TERMS USED

The limitations of the search interface of the HNC only allowed us to run a small variety of searches. Whereas in speech, focus is distinguished from a neutral constituent by stress and pitch accent, it is not marked directly in written Hungarian. Therefore, characteristic word order properties of focus were employed to retrieve a set of relevant corpus examples.

Focus shares these word order characteristics with certain other constructions, but they could nevertheless be used to extract examples automatically, assuming a manual filtering of the search results. In particular, the following two types of search were carried out, according to certain characteristics of Hungarian focus:

II.1. Searches for a sequence of “verb + verbal prefix”

The Hungarian focus position is immediately preverbal, and if the finite verb of the sentence is a prefixed verb, the verbal prefix is separated from the verb stem and appears to the right of the verb:
(1*) *János kinyitotta az ajtót.*
John pref-opened the door
’John opened the door.’

(2*) *János nyitotta ki az ajtót.*
John opened pref the door
’It was John who opened the door.’

(The number of examples which have not been taken from our database of examples has been marked with a star.)
Since verbal prefixes are POS tagged as such in the HNC, it was possible to search for sequences of finite verb plus verbal prefix, which is a search term that results in a large number of relevant hits. However, the postposing of the verbal prefix (put descriptively) is a property of several other syntactic constructions in Hungarian beside focus, especially negation of the finite verb, *wh*-questions (in which the question word appears in the focus position) and imperatives, e.g. for an imperative:
(3*) *Nyisd ki az ajtót!*
Open-imper pref the door
‘Open the door!’

Such structures and further irrelevant data (like examples from earlier historical stages of Hungarian, not straightforwardly interpretable texts like modern poetry, irrelevant hits due to mistakes in tagging) were manually discarded, which leaves about 40% of relevant examples.

II.2. Searches for a sequence of “noun (non-nominative) + verb”

Neutral word order in Hungarian is typically: subject first (as topic), followed by the finite verb, followed by the object and other complements of the verb. (Though it is not strictly necessary for the subject of the sentence to be the topic, and thus appear at the left edge of the sentence, this is at least the most usual combination of roles.) Whenever another order appears, i.e. the verb is preceded by a constituent other than the subject, it is quite likely that the preverbal constituent will be a focus. Thus this word order property was also used to search for foci.
Since functional roles of constituents (i.e. subject, direct object, beneficiary etc.) are relatively unambiguously marked by case morphology, and this is tagged mostly correctly in the corpus, it was possible to search for such sequences of a noun with a certain case ending (other than nominative) plus a verb. Some cases, in particular the one with the suffix *–ról* (meaning mostly ‘down from’ or ‘about’) and the one with the suffix *–n* (meaning ‘on’) yield an above average proportion of relevant examples (i.e. focus constructions), approximately 50% and 35% respectively. If we consider all searches of this type, i.e. for preverbal nouns with every available non-nominative case, the average proportion of focus examples in the search results was just around 25%. (These numbers all reflect what remains if clear cases of sentences with stress-avoiding verbs, complex predicate formation with a preverbal noun, and other supposedly irrelevant structures have been discarded from the results.)

As in the earlier case, searching for a sequence of a noun in other than nominative case and of a verb has the drawback of coinciding with constructions that are arguably independent of focus. These are, most significantly, instances of complex verb formation (by combining a non-nominative bare noun plus a finite verb, to yield a complex verb which behaves similarly to a prefixed verb) and uses of so-called stress-avoiding verbs (verbs which obligatorily expect a certain complement of theirs to appear in the focus position, without producing the interpretational effects characteristic of focus). We will return to this issue in detail below.
On the other hand, these searches have an advantage over searching for a verb plus a prefix, namely, that they retrieve focus examples both with verbs that are prefixed and ones that are not.
(4) *Hallgasson meg mindent, amit az orvos mond, de senkinek ne higgye el, amit [$_F$ a javulás várható esélyeiről] mondanak.*
'Listen to everything that the doctor says, but don't believe anyone what they say [$_F$ about the expected chances of recovery.]'

(Square brackets with a subscript F mark the focused constituent.)
In this example, the finite verb of the sentence is the simple (non-prefixed) verb *mond* 'say', which is preceded by a focused constituent with the typical contrastive identificational effect: you can trust in general what a doctor tells you, but you should not believe what he says about the probability of recovery.

Thus, both kinds of search term had certain advantages and drawbacks: searching for prefixes following the verb yields a relatively large number of relevant hits, but limits the range of verbal predicates under consideration to prefixed ones, whereas searching for non-subject nouns directly preceding the finite verb overcomes this latter limitation, but the ratio of relevant hits is relatively low, and separating relevant from irrelevant examples is often much more difficult. Therefore, we used both search methods extensively when compiling our examples database.

II.3. Searching for Hungarian clefts

There was one important further type of search, namely, for Hungarian cleft constructions. Intuitively, Hungarian clefts seem closer to English clefts in many cases both in their effects and the possible contexts of their use than Hungarian foci. An example of the Hungarian cleft construction is the following:

(5) *Az amerikai elnök nagyon reméli, hogy még a novemberi elnökválasztás előtt ő lesz az, aki áldását adhatja a közel-keleti békére.*
'The president of the United States is hopeful that it will be him, before the presidential elections in November, that can give the peace in the Middle East his blessing.'

The only practical way to search for Hungarian clefts, considering the limitations of the search interface, was to search for a sequence of the demonstrative pronoun *az* and a comma, which always appears in clefts. On average, about 20% of the hits of the results of such searches turned out to be clefts, which was a high enough number to work with effectively.

III. MANUAL PROCESSING OF SEARCH RESULTS

III.1. Discarding irrelevant search results

The results of searches in the HNC (a sample of around three thousand hits in total) were examined by Gergely Pethő, a native speaker of Hungarian. Completely irrelevant hits, such as sentences from historical texts, non-sentences (like titles), clearly uninterpretable or strongly ungrammatical sentences were discarded. Examples involving the postposing of the verbal prefix due to negation, imperatives and wh-questions (cf. example (3)) were mostly also removed, although we kept some of these to serve as illustration for these constructions.

III.2. Classification of search results according to descriptive groups

The search results which remained after these steps were included into our database of examples. Importantly, examples exhibiting the following phenomena were accepted into the database, even though generative approaches to the syntax of Hungarian traditionally assume that they are essentially independent of the focus construction per se, but only share some of its structural characteristics:

- Stress-avoiding verbs
- Complex predicate formation by preverbal bare nouns
- Focus-sensitive particles
- Quantifiers which have to appear in the focus position

We will discuss each of these groups in more detail in the following. The reason for not excluding them from the database out of hand directly followed from the aims of the research project: These constructions all involve a nominal element occupying the so-called focus position of the Hungarian sentence, i.e. the immediately preverbal surface position. The reason why they are considered to be independent of focus has to do primarily with semantics: it is normally assumed that focus has a particular semantics (exhaustive identification etc.), and these constructions differ in their interpretation from “true” focus. Since one of the main goals of our project was to explore the possibility that the traditional assumptions about the semantics of focus are incorrect (more exactly, too strong), it would have been illogical to exclude these examples on the grounds that they do not conform to usual ideas about what Hungarian focus is supposed to mean.

III.2.1. Stress-avoiding verbs

This category of verbs was introduced as a descriptive category by Komlósy (e.g. 1989). He observed that, as opposed to “normal” Hungarian verbs, some require the preverbal position to be filled whenever they are used, and thus force one of their arguments to appear in that position. This is the neutral structure for these verbs, and no special interpretation that would otherwise indicate a focus construction is associated with this word order. Compare

(6) *Másrészt nap nap után szembesülök azzal: megterhelt a hat év, amelyet [$_F$ a csapatnál] töltöttem.*
On the other hand, I am faced day by day with the fact that the six years that I spent [$_F$ with the team] have exhausted me.

State verbs such as *tölt* ‘spend time with’, like many verbs of movement, are stress-avoiding (although they often have a prefixed, non-stress-avoiding version as well). The property of being stress-avoiding does seem to be connected to some semantic and morphological aspects in general, but it does not seem possible to state rules that would allow one to predict whether any given verb of Hungarian does have this property.

Deciding whether a given sentence in Hungarian involves a true focus or a stress-avoiding verb requires one to consider alternative word orders with the same verb and decide whether those would be grammatical. To confirm that the example above does indeed contain a stress-

avoiding verb, one would have to check whether the neutral word order with unfilled focus position is grammatical:
(7) *…, *amelyet töltöttem a csapatnál.*
If a neutral word order is impossible in general, as in this case, one can conclude that the verb is in fact stress-avoiding. Note also that an alternative structure with the same locative complement, but the prefixed, non-stress-avoiding version of the same verb is indeed grammatical, which confirms that it is not e.g. some pragmatic reason that makes it necessary for the complement to be a focus, thus making (7) unacceptable:
(8*) *…, amelyet eltöltöttem a csapatnál.*

Note that whereas the focus-like word order of stress-avoiding verbs receives a neutral interpretation, it is possible to use these verbs with the contrastive interpretation that is characteristic of focus. In this case, the same word order is employed as in (6), but the preverbal constituent receives stronger stress:

(9) *A múlt hét óta már a Gazdasági Minisztériumhoz tartoznak a foglalkoztatáspolitikai feladatok.*
'Since last week, employment policy issues already belong to the Ministry of Economy.'

The verb *tartozik* 'belong somewhere, to something' is a stress-avoiding verb, the complement naming the place or entity to which the subject belongs has to appear in the focus position in a neutral structure. However, in (9), the Ministry of Economy is contrasted with the ministry that the employment policy issues belonged to before last week, thus serving as a true contrastive focus.

III.2.2. Complex predicate formation by preverbal bare nouns

Bare nouns in Hungarian are in general restricted to the immediately preverbal position in the Hungarian sentence in neutral structures. The presumed reason for this is that bare nouns do not form independent syntactic constituents, but can only appear in a sentence if they are lexically incorporated into the finite verb of the sentence. Thus bare nouns behave similarly to verbal prefixes in Hungarian, and belong to the syntactic category of verbal modifiers (VM) together with the prefixes. Some combinations of bare nouns plus verb form a complex lexical item and are stored as such in the lexicon, e.g. *divatba jön* 'become fashionable, lit. come into fashion':

(10) *Vannak iskolák, amelyek gyorsan [$_{VM}$ divatba] jönnek, és gyorsan leáldoznak.*
'There are schools which come [$_{\mathrm{VM}}$ into fashion] quickly, and decline quickly.'

Other verbs are not combined with their verbal modifier lexically, but combination of the two elements is rather a syntactic process, e.g. the bare noun *kamatemelés* 'raising of interest rates' as a VM to the verb *készül* 'prepare, be about to':

(11) *Stanley Fischer szerint az amerikai jegybank szerepét betöltő Fed [$_{VM}$ kamatemelésre] készül, az euró pedig erősödhet.*
'According to Stanley Fischer, Fed, which plays the role of the national bank of the USA, is about to [$_{\mathrm{VM}}$ raise interest rates], and the euro might become stronger.'

Similarly to prefixes, VMs have to be postposed when a true focus is used in the sentence and takes over the directly preverbal position, e.g. compare (8) and (10):

(12*) *Idén [$_F$ a sárga szín] jött [$_{VM}$ divatba].*
'This year, the colour yellow became fashionable.'

Like in the case of stress-avoiding verbs, it is possible for an incorporated bare noun to function as a true focus, receiving a contrastive interpretation, i.e. its preverbal position can be motivated in such cases in both ways. Verbs with incorporated bare noun complements can be distinguished from stress-avoiding verbs by checking whether the same verb can occur with a neutral word order (i.e. with an unfilled focus position) if it takes e.g. a definite NP complement instead of the bare noun, e.g.

(13*) *A Fed készül a kamatemelésre.*
'The Fed is making preparations for the raising of the interest rates.'

III.2.3. Focus-sensitive particles

Like in English, there is a set of particles in Hungarian the interpretation of which depends on the focused expression with which they combine in the given sentence (this phenomenon is usually referred to in the literature as association with focus). In Hungarian, it is the element in the focus position with which such focus-sensitive particles associate. It is rather clear that in terms of interpretation, foci which are associated with a focus-sensitive particle (sometimes called bound foci) are somewhat different from the more usual examples of "free" foci. The interpretation of such focused constituent is essentially a function of the focus-sensitive particle (which is thus frequently referred to as focus operator in general literature on these issues, e.g. Rooth 1992), instead of showing the interpretive effects commonly attributed to the Hungarian free focus (exhaustive listing, presupposition etc.).
The most common focus particles in Hungarian are *csak* 'only' and its variants, which convey the same semantic and presuppositional effects, but differ in stylistic effects and implicatures, e.g. *csupán, mindössze, kizárólag.* However, some less common focus particles were also represented among our search results, e.g. *legfeljebb* 'at most, at best', *egyenesen* 'directly', or the intensifier *maga* 'him/herself':

(15) *Segítséget, tanácsot [$_F$ legfeljebb környezetétől] kérhet ami viszont tovább fokozza a bizonytalanságot.*
'He can ask his environment at best for help and advice, which further increases the uncertainty, however.'

In this example, the particle *legfeljebb* modifies the focused expression *környezetétől* 'of (people in) his environment', and the resulting interpretation is that there is nobody the person in question could ask for help, except people in his environment, but even they are unlikely to help him.

III.2.4. Quantifiers which appear in focus position

It is a well-known property of the left periphery of the Hungarian sentence that quantifiers normally appear in canonical positions left of the verb, and their surface order determines their relative semantic scope. It is assumed (e.g. by Szabolcsi 1997) that the semantic properties of the quantifier determine which designated quantifier position a given quantifier expression appears in. One set of quantifiers, which can be descriptively termed "restrictive", seem to be normally assigned to the focus position, since, like true foci, they trigger the postposing of verbal modifiers, receive main stress of the sentence, and are forced to appear postverbally if a "true" focus appears in the sentence, which takes over that position. It is a complex theoretical issue whether these quantifiers really appear in the focus position (as was claimed in earlier work on the Hungarian left periphery, e.g. É. Kiss 1986), or whether they belong to another preverbal syntactic position that is very similar in terms of properties to focus, but not identical to it. There are rather few clearly syntactic arguments for assuming that what is involved here is a position different from that of foci, so we included such examples as potentially relevant to a theory of Hungarian focus. Compare the following example of this category:

(16) *Idén Stilelibero című albumát nyolcvanhat koncerten mutatja be.*
'This year, he presents his album titled Stilelibero in 86 concerts.'

Numeral quantifiers which have an 'exactly n' interpretation, like 86 in this sentence, appear in the (presumed) focus position, whereas their counterparts with the same form, but the interpretation 'at least n', appear in the so-called distributive quantifier position.

III.2.5. Focus

In addition to these four groups of examples, obvious cases of focus were naturally also included in our database. Some relevant examples are cited below in IV.5.

III.2.6. Cross-classification

Although we did not presuppose any necessary theoretically relevant difference between the four above-mentioned groups and focus, we did in fact characterise each example in the database on the basis of these descriptive categories. Assignment of individual examples to more than one group was possible and in fact occurred in many cases. Thus if the main verb of the example was a stress-avoiding verb, and in the given sentence its canonically preverbal complement was a bare noun (i.e. an incorporated verb modifier), and that bare noun would be used in the given context contrastively (which is a well-known interpretation possibility for bare nouns), then the given example would be assigned to all three descriptive categories.

IV. CONTENT OF THE DATABASE ENTRIES

For each example, an individual entry was created in the database. Information was entered into the database in the following database fields:

- Example
- Translation
- Descriptive group
- Search term
- Focus type

IV.1. Example

The field 'Example' contains a sentence in context that was returned as a result of one of the two types of search mentioned in section II. The central sentence always contains the verb-prefix or noun-verb sentence searched for, along with the annotation tags for the two expressions that reveal clearly why a given expression was considered a hit for the search terms by the search engine. The context that the sentences appear in is one of two or three sentences to the left and right of the sentence containing the hits. Including context is considered essential to judge especially what use focus is put to in the given case, e.g. whether it is contrastive. If the context provided in the database entry is still insufficient to decide why a focus was used, the source of the given example is marked in the HNC, so a greater context can easily be checked (although we found this only necessary in very few cases).

IV.2. Translation

This field contains an approximate English translation of the Hungarian example, or more specifically, usually the sentence that contains the hit plus, if necessary for correct interpretation, a further sentence, usually preceding this sentence.

IV.3. Descriptive group

This field specifies whether the part of the example that was returned as a hit for the search can be assigned to any of the alternative descriptive categories mentioned above (negation, stress-avoiding verb, focus-sensitive particle etc.) or, if no such alternative category can be identified, that it is a focus. Unclear status of the example was also specifically marked. In cases when an example contained several occurrences of the use of the focus position (e.g. in more than one part of a coordinate sentence), only a single such structure was characterised, namely, the one that was returned as the hit for the given search. For example, (3) contains an imperative in the main clause of the first conjunct and negation in the main close of the second one, but these categories were not included when classifying this example, because we were only interested in the focus of the subordinate clause of the second conjunct (bracketed in the example), which was the result of our corpus search (for a noun with the case ending -*ről* plus a finite verb).

IV.4. Search term

'Search term' specifies what search returned the given example as a hit. The category designations are used as in the search engine of the HNC. For example "*-bAn* (főnév) /+w1 ige" means a noun (főnév) with the case ending *-ban* or *-ben* (which is called inessive case in descriptive grammars of Hungarian), directly followed by (/+w1, using the COSMAS search engine notation) a verb (ige).

IV.5. Focus type

'Focus type' contains a preliminary characterisation of the function of the focus in those examples that we classified as 'focus' in the field 'descriptive group'. The categories used were 'identificational', 'contrastive' (which is generally regarded in the literature as an important subtype of the identificational use of focus), 'emphatic' (where focus appears to convey no identification and involve no presupposition, but only express emphasis), and 'other' for hard to classify cases.
Typical examples for each are the following:

IV.5.1. Identificational

(17) *Kigyulladt egy autóbusz városunkban a választások napján, azután, hogy már napok óta füstfelhőket eregetve közlekedett. A jármű [$_F$ a város központjában] gyulladt ki rövidzárlat következtében.*
'A bus caught fire in our town on the day of the elections. Earlier it had driven around the down for days spitting out clouds of smoke. The vehicle caught fire [$_F$ in the city centre] because of a short-circuit.'

Here, it is obviously known that the vehicle caught fire, and the author expects the readers to be interested in where and why this happened. She answers the first of these two questions that can be seen as implied by the context by a focus construction. This focus shows the usual interpretational effects that are attributed to identificational focus. Note that the answer to the question why the fire happened is also included in the sentence. It appears as a postverbal “information focus” according to É. Kiss (1998)’s terminology, i.e. as a non-presupposed, stressed postverbal constituent that introduces discourse-new information.

IV.5.2. Contrastive

(18) *Cáfolta azt, hogy Torgyán József vagy bármely más politikus közbenjárt volna Szenes kinevezése érdekében: Szabó [$_F$ szakmai meggyőződésből] jelölte e posztra.*
‘He denied that T. J. or any other politician had influenced the appointment of Szenes to her favour: she was appointed by Szabó for this position [$_F$ out of professional conviction].’

The author presumably wishes to contrast two possible reasons for the appointment of Szenes for the position under discussion: that a politician convinced Szabó to appoint her, or that Szabó made this decision because he thought that Szenes was, in terms of professional qualities, the best candidate for that position. The use of focus to express this contrast is quite straightforward.

IV.5.3. Emphatic

(19) *Ma az a megtiszteltetés ért, hogy [$_F$ Clinton elnök úr asztalánál] ebédeltem, 10-en ültünk az asztal körül, válogatottan olyan országok, amelyek az előkészítésben fontos szerepet játszottak.*
‘Today, I had the honour to eat lunch [$_F$ at the table of President Clinton], and we were 10 around the table; a select group of countries which played a significant role in the preparation process.’

In this case, focus does not seem to be used in the usual way: it would be odd to assume that the speaker considers it presupposed (and relevant to the hearers) that he ate lunch, the question being only where or with whom he ate. What focus seems to do in this case is to emphasise the fact that there was a lunch with Clinton, simply because it is automatically a noteworthy fact that one is honoured by the president of the United States. Similar examples were classified as emphatic.

IV.5.4. Other

(20) *Néhány török esett csak el, mivel azok szinte egyszerre mindnyáján [$_F$ a várba] kezdtek rohanni, ahova azután be is zárkóztak.*
‘Only some Turks died, because those started to run [$_F$ into the castle] all almost at the same time, where they later barricaded themselves.’

It is not completely clear what the reason for using a focus is in this case. The most likely explanation for it is that *mivel* ‘because’ associates with this focus, i.e. the reason for the Turks’ death is thought by the author to be the fact that they ran for the castle instead of in some other direction. A presupposition does not seem to be connected to the use of focus in this particular case.

V. SIZE AND CONTENT OF THE DATABASE

Our database of examples contains more than 1000 examples (of about 70,000 words in total), about 400 of which are examples of the use of the focus construction, the remaining 600 being mostly cases of the similar descriptive categories mentioned above. An important exception is the Hungarian cleft constructions, of which 100 examples were included in order to allow comparison with both Hungarian focus constructions and English clefts. These examples were not characterised further, apart from specifying ‘cleft’ as their descriptive group.

About 300 of the examples have been translated. Only the examples that we considered directly relevant for the research project received a translation exhaustively.