Sigurd, Bengt. 1985. 'Dialog på dator'. *Praktisk Lingvistik* 11, 33-52. Lund: Dept. of Linguistics

Snow, Catherine. 1979. 'Conversations with Children'. *Language Acquisition*, eds. Paul Fletcher and Michael Garman, 363-76. Cambridge: University Press.

Söderbergh, Ragnhild. 1980. 'En modell för beskrivning av dialoger mellan barn och vuxna' (summary in English: 'A model for the description of dialogues between children and adults'). *Språkstimulering i förskoleåldern*. Stockholm University: Child Language Research Institute, Paper 5.

Söderbergh, Ragnhild and Anne-Christine Bredvad-Jensen. 1987. 'A Dialogue Model and its Application in a Free Play Situation'. *The Nordic Languages and Modern Linguistics* 6, 369-396. Helsinki: Helsinki University Press.

Wardhaugh, Ronald. 1985. *How Conversation Works*. Oxford: Basil Blackwell.

Lund University, Dept. of Linguistics
Working Papers 33 (1988), 91-102

Preliminaries to a Referent Grammatical Analysis of Modern Irish Relative Clauses

Sheila Dooley Collberg

Abstract

Following is a report on work in progress on the development of a computer analysis of Modern Irish for use in machine translation and generation. The program uses Referent Grammar (RG) and gives special attention to the analysis of relative clauses. The program is still in a very early stage of development, so discussion is limited to a survey of the technical problems involved in the construction of a referent grammatical analysis for Irish. Some of these problems include inherent structural ambiguities, proper morphological treatment of inflected prepositions, and eventual parsing of complex relatives. The possibility of generating syntactic parse trees along with the usual functional representation output of RG is also considered as a desirable expansion.

BACKGROUND

Referent Grammar (RG) is a GPSG-based formalism designed for use in computer analysis of language. RG has been used successfully in systems for generation and translation of a variety of languages – Swedish, English, French, Georgian, Polish, and Samoan – in conjunction with work done by SWETRA (Swedish Computer Translation Group) at the University of Lund, Sweden. The preliminary analysis of Irish presented here is modelled on these existing modules and is intended to be expanded for use with them. Since Irish is a VSO language of the Celtic family, its analysis forms an important typological expansion of SWETRA's capacities. Irish also presents some interesting technical problems for the formalism of RG to describe. These will be discussed after a short introduction to the basics of RG.

REFERENT GRAMMAR

The fundamentals of RG are explained fully in Sigurd 1987 and 1988, and the Irish analysis presented here relies heavily on the analysis of Polish relative clauses given in Gawrońska-Werngren 1988. The reader is referred to these articles for detailed explanations. I will, however, briefly review the features which characterize RG and which are essential to an understanding of this paper.

First, an analysis written in RG is written directly in the Prolog programming language using Definite Clause Grammar (DCG) formalism. Thus, the syntactic rules of RG look like the kind of syntactic rules one usually encounters in generative grammars, with a mother constituent to the left of an arrow and its daughter constituents listed to the right. Unlike most generative grammatical rules, however, RG rules make use of a functional representation (f-rep) in addition to a linear and hierarchical syntactic representation (o-rep). The inventory of functional categories which may appear in an f-rep has been "standardized" in order to serve as an interface for translation between different languages. Example (1) gives the basic form for a sentence rule in RG (a) and shows the standardized form for f-reps (b).

(1) a. sent(M,P,X,F)→. . . ; where:
M=mode
P=an undefined variable which may be used for polarity
X=the focused constituent of the sentence
F=the functional representation
b.s(subj(X),padv(AO),pred(Y),dobj(Z),obj(O),sadvl(S1),sadvl(S2),
advl(A1),advl(A2),advl(A3); where:
padv=particle clause
sadvl=sentence adverbial(negatives, etc.)
dobj=dative (indirect) object

RG is GPSG-based in that it makes use of defective constituents (usually sentences) for parsing. Relative clauses, for example, are effectively described as defective clauses: they lack some constituent which may be identified with the head noun of the relative construction. This will be explained in detail in the following sections.

Finally, the most important aspect of RG is that it uses referent variables. That is, each referent introduced into discourse receives a number – a referent variable – which may be matched with instances of the same referent in other sentences or in the same sentence. RG in this way provides a formal method for the indexing of noun or pronoun referents which other theoretical frameworks cannot always describe clearly. In addition to bearing a referent number, each referent carries with it an array of grammatical and semantic information which is essential for effecting complete and correct translations between languages. Example (2) shows the standard form for encoding this array of information, which is known as the "enriched" referent.

(2)r(R,A,D,N,S,G1,G2)
where R=referent number
A=meaning, given in "machinese" English
D=definiteness
N=number
S=sex
G1=grammatical gender (in the language to be translated to)
G2=grammatical gender (in the language translated from)

ITRANS

The Irish program which will be presented here has been given the name ITRANS. In its current state, ITRANS is capable of parsing both whole sentences and noun phrases in isolation. However, it is as yet restricted to handling only certain types of main sentences, namely those with the basic word order configurations shown in (3). It was deemed best to concentrate on these basic types in order to work out the intricacies of a preliminary referent grammatical analysis. It was also necessary to develop an analysis for these basic sentence types since relative clauses will be described as defective variants of them. Eventually ITRANS will be expanded to handle a broader range of sentence types.

(3) a.VS: Léigh fear.
read man 'A man read.'
b.VSO: Molann an fear Bríd.
praise the man Bridget 'The man praises Bridget.'
c.VSOp[O]: Thug an fear cat do Bríd.
gave the man cat to Bridget 'The man gave a cat to B.'

Irish Relative Clauses

Relative clauses have been described as a combination of a head noun which specifies a domain and a subordinate clause which restricts that domain to a particular subset (Keenan and Comrie 1977). In Irish, relative clauses are postnominal modifiers which appear inside noun phrases. ITRANS therefore gives particularly close attention to the analysis of noun phrases. The syntactic structure of Irish noun phrases used by ITRANS has been adapted from McCloskey 1979. The noun is the only obligatory constituent; all others are optional. The determiner position is only filled in the case of definite nouns, and any adjectives must precede any relative clauses.

The syntax of modern Irish relative clauses has been well documented by McCloskey 1979 and 1985 and Harlow 1981. Irish uses a mixed system of relativization with gapping restricted to certain constructions and resumptive pronouns required in others. These two types of relatives are traditionally referred to as the "direct" relative (with a gap) and the "indirect" relative (with a resumptive pronoun). They are distinguishable from each other not only in the presence or absence of a resumptive pronoun but also in the nature of the phonetic mutation accompanying the relative marker. The relative marker in Irish always takes the form *a*; however, this *a* induces lenition on the following verb in a direct relative construction, while it induces eclipsis on the following verb in an indirect construction.[1] For the sake of clarity, I will follow the usual practice in rendering the leniting marker with *aL* and the eclipsing one with *aN*.

The direct relative is obligatory when the relativized head noun functions as the subject of the relative clause. This is illustrated in (4), where the site of the gap is marked by a blank space.

(4) an fear aL dhíol _ an bád
the man REL sold _ the boat

Most other relative constructions – relativized prepositional objects, genitives, and objects of comparison – are formed with the indirect relative. Examples of each type are given below in (5-8). The direct relative is always ungrammatical with these constructions.

(5) an fear aN dtabharann tú an t-airgead dó
the man REL give you the money to-him
'the man to whom you give the money'

(6) an fear aN bhfuil a mháthair sa bhaile
the man REL is his mother at home
'the man whose mother is at home'

(7) fear aN bhfuil mé níos mo na é
man REL is I bigger than him
'a man whom I am bigger than'

[1] Lenition and eclipsis are systematic phonological changes which function in the Celtic languages to signal a variety of morphological prcesses. Lenition is the process by which consonants become more lenis in articulation: stops become fricatives. The fricatives [f] and [s] and the nasal [m] are also affected. Eclipsis (also known as nasalization) is the process by which voiced segments become nasalized and voiceless segments become voiced.

The only instance in which the choice between using a direct or indirect relative is optional is when the relativized noun referent functions as the direct object in the relative clause. The two types of Irish relatives thus overlap partially in the relativized direct object construction. Examples of the direct and indirect variants are given in (8).

(8) a.an scríbhneoir aL mholann na micléinn _
the writer REL praise the students _
'the writer whom the students praise'
b.an scríbhneoir aN molann na micléinn é
the writer REL praise the students him

RG treats all relative clauses as defective variants of subordinate sentences, or "sunts". Those containing a gap are fully defective since they lack one constituent in surface form, while clauses containing a resumptive pronoun may be described as "semi-defect" in order to indicate that they are true relatives in function yet lack no constituents in surface form. ITRANS thus defines the Irish direct relatives as defective and the indirect relatives as semi-defective. The name given to each type of relative clause reflects which constituent has been relativized and whether the clause is fully defect or semi-defect:

(9) Example (4) sdsunt = subject-defect subordinate sentence
(5) semipodsunt = semi-prepositional-object-defect sunt
(6) semigdsunt = semi-genitive-defect sunt
(7) semicomdsunt= semi-comparitive-defect sunt
(8a) odsunt = object-defect sunt
(8b) semiodsunt = semi-object-defect sunt

The functional representations of ITRANS ensure that the missing constituent of the relative clause is identified with the head referent of the entire noun phrase and that the relative clause consequently receives the same referent number as the head noun. Since the rules of ITRANS distinguish between defect and semi-defect relative clauses, it is relatively easy to control the appearance of resumptive pronouns and of the relative marker so as to prevent, for example, *aN* from being used with an sdsunt or odsunt. Just as ITRANS does not yet include rules for all main sentence types, it is not yet equipped to handle all relative clause types illustrated here. Presently ITRANS includes rules for sdsunt, odsunt, semiodsunt, and semipodsunt. The remainder of this paper will discuss problems involved in implementing these rules and rules for the remaining types of relative clauses in a fully-developed translation program.

Ambiguities

Semantic and syntactic ambiguities are far from rare in natural language. They arise frequently, but they can usually be resolved by using clues from context or intonation. A closer look at the Irish direct relative reveals that its optional use for relativized direct objects can give rise to syntactic ambiguity. The ambiguity originates in the rigid VSO word order of Irish and lack of nominative/accusative case distinctions on Irish nouns. Compare again the direct relatives for subjects and direct objects, repeated here from examples (4) and (8):

(10) a. an fear aL dhíol _ an bád
b.an scríbhneoir aL mholann na micléinn _

Notice that the surface word order inside the relative clause in both examples is verb + noun. In (10a) we are dealing with a subject defect clause, while in (10b) we are concerned with an object defect clause. Nevertheless, this difference is in no way signalled by the surface order of the constituents. It can be seen here because we have marked the site of each gap by a blank space, but this marking is of course only a formal device and has no reality whatsoever in the language itself. The most obvious way to convey the fact that *bád* in (10a) should be interpreted as an object and *micléinn* in (10b) as a subject would be to mark them morphologically in some way, for example with morphological nominative and accusative case. But as stated earlier, Modern Irish no longer marks any nominative/accusative case distinctions. Genitive and vocative case are the only cases which survive in separate forms, and these are usually produced by changes in consonant quality rather than by suffixation. Nominative/accusative case is referred to simply as "common" case and is, as its name indicates, the common case for both subjects and objects. Given this, there is ample justification for interpreting example (10b) in a totally different manner, i.e. with *micléinn* functioning as object and the relative clause labelled as an sdsunt. This interpretation, given in (10b'), is in fact acceptable.

(10b') an scríbhneoir aL _ mholann na micléinn
'the writer who praises the students'

Not even the form of the verb inside the relative clause can offer any clue to the identity of the subject here. While Irish verbs may be inflected for person and number of the subject, markings are never present when there is an overt subject or a trace filling the subject position (McCloskey and Hale 1984). How, then, do Irish speakers disambiguate the direct subject and object relatives? And, more importantly, how should a program for machine translation of natural language deal with this inherent structural ambiguity? It has already been shown that Irish speakers do not have recourse to the most common grammatical means of signalling functional structures: word order and inflections on nouns and verbs. Intonation cannot give any clue either. The phonetic interpretations of pairs like (10b) and (10b') are as far as can be seen identical. Furthermore, any phonetic cues which might be found are useless in written texts, where these ambiguous examples appear fairly frequently.

If we try to construct an alternate interpretation for example (10a), however, it becomes evident that semantics may be able to provide some help:

(10a') an fear aL dhíol an bád _
? the man whom the boat sold

(10a') is semantically strange since we would not expect an inanimate object such as a boat to be acting as a seller of an animate object, especially one that is human. This might lead us to describe a strategy in which Irish speakers use semantic knowledge and expectations about the world to assign agent and patient roles to noun referents. Incorporating such a strategy into a computer translation facility would of course require describing an intricate and sophisticated database of semantic relations and rules for making inferences from the information stored there. But even access to such an elaborate database could not resolve all the ambiguities which might arise, as a simple look back at (10b-b') will confirm. In instances where the given noun referents are semantic "equals" and both fully capable of assuming the role of agent, there is still no way to choose one interpretation over the other. Writers and students are capable of mutual praise. Finally, semantic strategies based on expectations are always open to failure since expectations may not always be fulfilled. We expect cats to chase rats, but it is conceivable that we might need to describe a situation somewhere in which the rat would be doing the chasing.

While a semantic strategy does not provide a satisfactory answer to the problem of how Irish speakers deal with the subject/object ambiguity in relative clauses, there is indication that they do use what is best described as a default parsing strategy. When confronted with a potentially ambiguous relative, native Irish speakers will always choose to interpret it as a relativized subject, and they may even have to be prompted to confirm that an object interpretation is equally plausible (James McCloskey, personal communication). This default parsing strategy combined with the use of context seems to be the key to native speakers' tolerance of subject/object ambiguities.

Although context is much too elusive a concept to be used in machine translation, a default parsing strategy may be exploited efficiently. This is in fact the method which other referent grammatical modules have used for dealing with ambiguity. For example, a similar ambiguity between subject and object is also possible in Swedish, which, like Irish, lacks case distinctions on nouns:[2]

(11) Råttan jagade katten
rat-the chased cat-the

(11) may either mean 'The rat chased the cat', or 'The rat, the cat chased' (where *råttan* has been topicalized). In the Swedish RG module, the subject interpretation is defined as the default case and is the one which is always used as the basis for translation.

The current version of ITRANS follows this method and defines the subject interpretation of all direct relatives as the default case. While this is an accurate reflection of native speaker competence in dealing with sentences out of context, it is probably not as accurate for sentences in context. The default strategy in effect restricts the use of the direct relative to relativized subjects and does not recognize the optionality allowed with relativized objects. It totally ignores the fact that the direct relative is actually the preferred option for forming relativized direct objects (McCloskey 1985). The textual frequency of direct relatives used with objects demands that a system for translation be able to recognize and process them correctly. It is therefore desirable that ITRANS should have access to both interpretations, and work is still in progress on achieving this goal. The default strategy may still be used, giving the subject interpretation first in every case. But the object interpretation should also be returned by the system as a possible alternant. The choice between the two alternates would then be an editorial one to be made by a human editor capable of interpreting context.

Inflected prepositions

Notice that prepositional relatives such as the one which was illustrated in (5) are analyzed by ITRANS as semi-defect clauses, yet (5) seems to lack a resumptive pronoun. Such examples do contain a resumptive pronoun, however, incorporated into the preposition *dó*. As the translation indicates, *dó* may be

[2] I would like to thank Bengt Sigurd for pointing this fact out to me.

described as an amalgamated form of *do* + *é* (= 'to him'). The sixteen most common Irish prepositions systematically merge with pronouns to produce "inflected prepositions". The full paradigm for the preposition *do* is given below to illustrate this.

(12) do + mé = dom	'to me'
tú = duit	'to you'
é, í = dó, di	'to him, her'
sinn = dúinn	'to us'
sibh = daoibh	'to you (pl)'
iad = dóibh	'to them'

How should the rules of ITRANS relate the surface form of inflected preposition – one entity – to the underlying form of prepostion + resumptive pronoun – two entities? One alternative, of course, would be to abandon the idea of inflected prepositions altogether and describe forms of *do* from the opposite perspective: i.e., that *dom, duit, do*, etc. are inflected forms of the personal pronouns (call them "dative", if you will). The appearance of these pronouns can then be left to agreement restrictions on the forms of pronouns allowed in particular contexts. While this approach might seem to make for a simpler program, it is hardly justifiable in view of the case we have just seen for interpreting *dom, duit*, etc. as inflected prepositions rather than as inflected pronouns. Recall that Irish only distinguishes between common, genitive, and vocative case. It is not reasonable to argue for describing a whole dative case on the basis of pronoun forms which clearly include some form of preposition. Nouns never show the same behavior when preceded by a preposition. But this approach is not only descriptively inappropriate; it also proves to be practically inefficient. Defining inflected prepositions as casemarked pronouns does not simply mean defining a "dative" case for each pronoun. It actually involves defining at least sixteen different forms for each pronoun and rules for choosing from among them.

Interpreting *dó* as an inflected preposition is by far a much better course of action, even though it still involves giving separate lexicon entries for each inflected form. The ideal solution, both from a technical and descriptive standpoint, would be to define systematic rules for the formation of the inflected forms and let ITRANS use these in its analysis. But there is a good deal of irregularity in the ways in which the prepositions and pronouns combine, so truly systematic rules are not possible. Instead, ITRANS treats the inflected prepositions as a separate word class which just happens to have a meaning consisting of two elements. The meanings of both the preposition and its incorporated pronoun are both given in the meaning slot of the lexicon entry, and they may be called there and returned separately in their appropriate places in the functional representation built by ITRANS.

Parse trees

In the existing versions of RG, sentence analysis consists strictly of the functional representation of the sentence with the meaning and other grammatical properties (definiteness, number, sex, etc.) associated with the enriched referent. This is understandable since these are the factors which are essential to translation. However, RG analysis still involves parsing sentences at the word level in order to determine whether they are grammatically acceptable, and there is no reason to simply discard this syntactic parse once it has been effected. It can be printed out in addition to the functional representation returned with the translation information. Furthermore, it can be used as input to the Showterm package available for use with Prolog systems in order to generate computer-drawn parse tree diagrams. Although it is possible to return such syntactic parse trees along with functional representations, it does demand a great deal of extra control over the program. The syntactic parse must appear as an extra argument in each syntactic rule of ITRANS. This means more variables to keep track of, and more chance for conflicts to arise between form and function. Inflected prepositions, for example, are an obvious trouble spot in this respect. Nevertheless, parse trees are a desirable addition to ITRANS, especially when one considers the inherent ambiguities which have already been demonstrated concerning subject and object relatives. A syntactic parse drawn out in tree form can serve as an extra control and a separate aid for a human editor who must decide whether the ambiguous sentences have indeed been interpreted correctly by the system.

Directions for further developments

The previous sections have already sketched some of the features which are currently being developed in ITRANS: access to multiple interpretations, correct syntactic and functional analysis of inflected prepositions, generation of syntactic parse trees as well as functional representations, and the expansion of the entire system to include rules for the full range of sentence and subordinate clause types in Irish. Some more refined additions may also be mentioned briefly here:

1. Other forms of the relative marker.

The relative markers *aL* and *aN* assume other shapes when the verb of the relative clause is negated or appears in past tense. ITRANS should include rules for restricting even these forms to appearing in the correct contexts. Most importantly, the system must recognize that negation appears incorporated into the relative marker and not as a preverbal particle as in main sentences.

2. Analytic and synthetic verb forms.

Irish verbs usually have two variants, an analytic one which shows only tense/mood information, and a synthetic one which encodes person-number inflection as well as tense/mood. The use of the synthetic forms in relative clauses is dependent upon the shape of the relative marker and upon the combinations of simple and complex relative constructions which are possible.

3. Complex relatives.

ITRANS is even now syntactically capable of parsing multiply embedded relatives or complex relatives. However, complex relatives in Irish do not follow precisely the same rules as simple relatives with respect to the use of the direct and indirect relative markers. There are intricacies of the *aL/aN* alternation to be described, with the complementizer *goN* also used as a relative marker on certain embedded relatives. See McCloskey 1985 and McCloskey and Hale 1984 for detailed discussions of this phenomenon. An example is given here for illustration.

(13) na daoine aN raibh mé ag dúil goN gcuirfidís isteach ar an phost sin
the people REL was I expecting REL put(COND-3pl) in on that job
'the people that I expected (that they) would apply for that job'

Notice here that the site of relativization is the subject of the embedded clause, yet the indirect marker *aN* and the complementizer *goN* are used. The inflection on the synthetic verb form of the conditional serves in this case as resumptive pronoun.

REFERENCES

Gawrońska-Werngren, Barbara. 1988. 'A referent grammatical analysis of relative clauses in Polish.' *Studia Linguistica* 42(1).

Harlow, Stephen. 1981. 'Government and relativisation in Celtic'. *Binding and filtering*, ed. Frank Heny, 213-254. London: Croom Helm.

Keenan, Edward and Bernard Comrie. 1977. 'Noun phrase accessibility and universal grammar'. *Linguistic Inquiry* 8, 63-99.

McCloskey, James. 1979. *Transformational syntax and model theoretic semantics: a case study in Modern Irish.* Dordrecht: D. Reidel.

McCloskey, James. 1985. 'The Modern Irish double relative and syntactic binding'. *Ériu* 36, 45-84.

McCloskey, James and Kenneth Hale. 1984. 'On the syntax of person-number inflection in Modern Irish'. *Natural Language and Linguistic Theory* 1, 487-533.

Sigurd, Bengt. 1987. 'Referent Grammar (RG) in computer comprehension, generation, and translation of text (SWETRA)'. *Working Papers* 31, 137-67. Lund: Dept of Linguistics.

Sigurd, Bengt. 1988 (forthcoming). 'A referent grammatical analysis of relative clauses'. *Studia Linguistica Hafniensia.*

Lund University, Dept. of Linguistics
Working Papers 33 (1988), 103-121

Some Types of Russian and Polish Interrogative Sentences
– A Preliminary Referent Grammatical Analysis

Barbara Gawrońska-Werngren

INTRODUCTION

In the preliminary analysis presented in this paper, referent grammatical rules of the type developed in Sigurd 1987 will be applied to some examples of Russian and Polish *yes/no* questions and *wh*-questions. Our aim is to investigate the possibility of computer translation between English and Swedish and the Slavic languages by means of RG – a theory that the computer parser used by SWETRA (Swedish Computer Translation Group, Lund) is based on.

As RG is a phrase structure grammar, inspired by GPSG (Gazdar, Klein, Pullum and Sag 1985), such constructions as interrogative sentences and relative clauses are not analyzed as results of transformations, but as structures containing a topicalized constituent and a defective sentence. This means that a question like *who hit him* is analyzed as (using a simplified notation):

```
esent(q,T,P,F) → enpqs(P), esdsent(_,_,P,s(subj(P),pred(B),obj(X))).
                  who                  who       who     hit    him
```

q = question
T = tense
P = the focused constituent
F = functional representation
esdsent = English subject defective sentence
enpqs = interrogative noun phrase (subjective form)

The questions introduced by a finite verb are analyzed as verb defective (or aux-defective); if the objective *wh*-word is fronted *(whom did he hit)*, the sentence following the interrogative pronoun is classified as object defective, and so on. The computer translation is based on the functional representation of a sentence,