Analytic and Synthetic Verb Forms in Irish – An Agreement-Based Implementation in LFG

Sebastian Sulger
Department of Linguistics
University of Konstanz
Germany
sebastian.sulger@uni-konstanz.de

Abstract

This paper discusses the phenomenon of analytic and synthetic verb forms in Modern Irish, which results in a widespread system of morphological blocking. I present data from Modern Irish, then briefly discuss two earlier theoretical approaches. I introduce an alternative, agreement-based solution, involving 1) a finite-state morphological analyzer for verb forms implemented using the FST toolset (Beesley and Karttunen, 2003); 2) a computational grammar of Modern Irish couched within Lexical-Functional Grammar (LFG) (Bresnan, 2001) and implemented using the XLE grammar development platform (Crouch et al., 2008).

1 Introduction

In Modern Irish, verbs may appear in two different forms: synthetic and analytic. Across tense and mood paradigms, certain person-number combinations are expressed by markers on the verb, resulting in so-called synthetic verb forms. Other person-number combinations are expressed by personal pronouns which appear independent of the verb.

(1) Present tense paradigm for *tuig* 'understand':

1P.Sg	*tuigim*	'I understand'
2P.Sg	*tuigeann tú*	'you understand'
3P.Sg.M	*tuigeann sé*	'he understands'
3P.Sg.F	*tuigeann sí*	'she understands'
1P.Pl	*tuigimid*	'we understand'
2P.Pl	*tuigeann sibh*	'you understand'
3P.Pl	*tuigeann siad*	'they understand'

In this example, the forms of the first person singular and the first person plural are synthetic forms. Person and number information is expressed by the ending on the verb. The other forms are analytic verbs which involve separate personal pronouns to express person and number information.

It has been acknowledged in the literature that the function of the person-number endings on the synthetic forms are identical to the function of the independent personal pronouns (Andrews, 1990; Legate, 1999). Evidence for this comes from two observations. First, the use of an independent personal pronoun is prohibited in conjunction with a synthetic verb form.

(2) *Tuigim mé an fhadhb.
understand.Pres.1P.Sg I ART problem
'I understand the problem.'

(2) is ungrammatical because the person-number information is realized on the synthetic verb form, blocking the use of a separate personal pronoun.

Second, the use of an analytic verb form is blocked if there is a synthetic verb form realizing the same features as the analytic form combined with a pronoun. Since there is a synthetic verb form available in the paradigm for the features '1st person singular' (*tuigim*), the use of the analytic verb form in conjunction with a personal pronoun is blocked.

(3) *Tuigeann mé an fhadhb.
understand.Pres I ART problem
'I understand the problem.'

An implementation using a computational grammar is thus faced with two separate tasks: 1) block the redundant use of the independent subject pronoun when combined with a synthetic verb form, as in (2); 2) block the analytic verb form when there is a synthetic verb form available, as in (3).

2 Earlier Approaches

Andrews (1990) presents an LFG approach. The approach crucially depends on the mechanism of

unification. More specifically, he proposes a solution in form of a constraint on lexical insertion, the *Morphological Blocking Principle*. Andrews (1990) defines this principle as a variant of the Elsewhere Condition, modified to control lexical insertion in LFG. The principle is formulated as follows:

Morphological Blocking Principle (MBP):

> Suppose the structure S has a preterminal node P occupied by a lexical item l_1, and there is another lexical item l_2 such that the f-structure determined by the lexical entry of l_1 properly subsumes that determined by the lexical entry of l_2, and that of l_2 subsumes the f-structure associated with P in S (the complete structure, after all unifications have been carried out). Then S is blocked.
>
> (Andrews, 1990, p. 519)

For Irish verbs, this principle essentially has the following consequences. When the f-structure of an analytic verb form is unified with the f-structure of an independent pronominal, the lexicon has to be checked to see if there is another verb form that subsumes the resulting unified f-structure (i.e., a form that already contains the pronominal features in its lexicon entry – a synthetic form). If there is such a form, the analytic form is blocked.

An obvious issue with this approach is connected to efficiency. For every verb form occurring in a sentence, the whole lexicon has to be checked for a corresponding synthetic form. While Andrews (1990) claims that a first small implementation by the author computes morphological blocking at a tolerable rate, it remains questionable whether this approach is adequate for larger-scale grammars.

Legate (1999) proposes a treatment of morphological blocking based on agreement. The analysis is couched within the framework of Distributed Morphology, drawing on insights from McCloskey & Hale (1984). It argues that the affixes found on verbs are truly agreement patterns in Modern Irish. The agreement between the verb and the subject pronoun must be realized via an agreeing affix on the verb (i.e. the synthetic form), since these affixes are more specified than the default affix (i.e. the analytic form). The paper departs from earlier literature in Distributed Morphology in that it requires two changes in the vocabulary insertion mechanism. First, the mechanism must operate top-down instead of bottom-up, as was assumed in previous papers. Second, any morpho-syntactic features realized by a vocabulary item have to be deleted. The paper concludes arguing that the Irish data constitutes an interesting argument for a framework of morphology that applies post-syntactically, based on competition.

Legate (1999) also mentions the paper by Andrews (1990), saying that this lexicalist alternative is problematic as it involves trans-derivational comparison (i.e., the MBP), which is a powerful and costly mechanism. Since Andrews (1990) compares the blocking of analytic forms by synthetic forms to the blocking of expressions like *the day before today* by *yesterday*, Legate (1999) concludes that a mechanism like the MBP eventually restricts wordiness.

To sum up, the paper by Andrews (1990) presents a first LFG account for the Irish data, but fails to provide an efficient implementation of the solution, although the approach is theoretically interesting. Legate (1999) makes convincing arguments for an agreement analysis, but, being a theoretical paper, does not offer an implementation; the paper also has to make changes to the applied theory of Distributed Morphology in crucial places.

3 An Alternative LFG Implementation

In this section, I present an alternative LFG approach to the problem of analytic and synthetic verb forms, drawing on theoretical insights from McCloskey & Hale (1984) and Legate (1999). I agree with their work in assuming an agreement relationship between the verb and the subject pronoun. Instead of a competition-based approach (Legate, 1999), my solution constitutes a lexicalist alternative based on agreement and unification, similar to the approaches by Butt (2007) for Punjabi and Bresnan (2001) for Navajo.

My solution uses agreement equations between the verb and the pronominal as a means to block analytic forms from occurring where synthetic forms are available. The implementation is twofold: 1) A detailed finite-state morphological analyzer (FSMA) dealing with Irish verbal morphology has been written, listing both analytic and synthetic verb forms with morphosyntactic features; 2) a computational LFG grammar has been implemented which effectively rules out analytic verb

forms in inappropriate contexts using two short agreement templates. The implementation is situated in the frame of the ParGram project (Butt et al., 2002). For a related implementation of Welsh morphology, the reader is referred to Mittendorf and Sadler (2006).

3.1 The Finite-State Morphological Analyzer

The FSMA I implemented lists both analytic and synthetic verb forms in the lexicon. All forms are provided with appropriate morphosyntactic tags. The FSMA was implemented using the FST toolkit (Beesley and Karttunen, 2003). In (4), I give the complete present tense paradigm for *tuig* 'understand' and the analysis for each of the verb forms.

(4) Pres. tense paradigm for *tuig* & FST analysis:

1P.Sg *tuigim*
`tuig+Verb+Pres+1P+Sg+PronIncl`
2P.Sg *tuigeann*
`tuig+Verb+Pres+2P+Sg`
3P.Sg.M *tuigeann*
`tuig+Verb+Pres+3P+Sg`
3P.Sg.F *tuigeann*
`tuig+Verb+Pres+3P+Sg`
1P.Pl *tuigimid*
`tuig+Verb+Pres+1P+Pl+PronIncl`
2P.Pl *tuigeann*
`tuig+Verb+Pres+2P+Pl`
3P.Pl *tuigeann*
`tuig+Verb+Pres+3P+Pl`

Notice two things about this analysis. First, the tag `+PronIncl` is attached to synthetic verb forms. This is to make sure that the subject receives a pronominal analysis and a PRED value – details follow in the next section.

Second, the forms are provided with detailed person and number information, even though the verb form is identical in some cases (i.e., the forms are not marked for person/number). A detailed analysis like this enables the grammar to enforce agreement constraints and effectively rule out analytic forms where synthetic forms are available – again, details follow in the next section.[1]

3.2 The Computational LFG Grammar

The grammar, implemented using the XLE grammar development platform (Crouch et al., 2008), makes use of the detailed morphosyntactic information provided by the FSMA. The grammar uses a template to enforce agreement restrictions, thereby ruling out analytic forms where synthetic forms are available.

First, I show how the grammar rules out the combination 'synthetic verb form + independent subject pronoun' (see (2) for an example). Recall that synthetic forms are provided with the tag `+PronIncl`. Associated with this tag is the following information in the tag lexicon of the grammar:

(5) Information associated with `+PronIncl`:
`PronSFX = (↑ SUBJ PRED) = 'pro'`

That is, the tag itself provides the information that the subject is a pronominal. In contrast, the lexicon entry of an independent pronoun is given in (6).

(6) `mé  PRON * (↑ PRED) = 'pro'`
`(↑ PRON-TYPE) = pers`
`(↑ PERS) = 1`
`(↑ NUM) = sg.`

When a pronoun like this occurs in the subject position after a synthetic verb form, the unification fails, since there are multiple PREDs – the one supplied by the synthetic form and the one supplied by the pronoun. Multiple PRED features for a single grammatical function are not allowed by LFG, since PRED features are not subject to unification (Bresnan, 2001; Butt, 2007).[2]

Second, I turn to the more difficult case: how to prevent analytic forms from occurring when synthetic forms are available (e.g., how to rule out sentences like (3)). Recall the detailed morphosyntactic analysis of verb forms outlined in section 3.1. Again, there is functional information associated with each of the tags in (4); see the entries in (7).

[1] One reviewer asks whether it would be possible to use a single non-1st-person feature instead of multiple feature sets for the same verb form. This is a question which largely depends on the application for which the FSMA was developed. While the features might not be strictly necessary as input to the LFG grammar to check for agreement facts, they might become a) important to check for in other places in the LFG grammar; b) important to check for by other applications which might be able to make use of the FSMA. Therefore, I have decided to keep the tags. A related discussion in connection with the German ParGram grammar is whether one should have morphological case tags for nouns which have the same form in all cases, where it was decided to include an `.NGDA` tag for such nouns (Butt, personal communication).

[2] One reviewer asks about how ungrammatical sentences such as **tuigeann __ an fhadhb.* are handled, where the verb form is not synthetic in nature and there is no subject pronoun. Sentences like these essentially violate the principle of completeness in LFG, stating that predicators must be satisfied by arguments with semantic features, i.e. PREDs. The above sentence therefore is ungrammatical since the verbal predicator demands a subject argument PRED, and since there is no subject, cannot be satisfied; see also Bresnan (2001).

(7)
```
+1P V-PERS_SFX XLE @(AGR-P 1).
+2P V-PERS_SFX XLE @(AGR-P 2).
+3P V-PERS_SFX XLE @(AGR-P 3).
+Sg V-NUM_SFX XLE @(AGR-N sg).
+Pl V-NUM_SFX XLE @(AGR-N pl).
```

These entries call up templates in the grammar, passing values over to them. For example, the entry for the morphological tag `+2P` calls up the template `AGR-P` and passes over the value `2`; similarly, the entry for `+Sg` calls up the template `AGR-N` and passes over the value `sg`. I provide the templates `AGR-P` and `AGR-N` in (8) and (9).

(8) `AGR-P(_P) = (↑ SUBJ PERS) = _P.`
(9) `AGR-N(_N) = (↑ SUBJ NUM) = _N.`

When the value `2` is passed on to `AGR-P`, the template tries to assign the value to the `PERS` attribute of the subject; correspondingly, when the value `sg` is passed on to `AGR-N`, the template tries to assign the value to the `NUM` attribute of the subject. The templates effectively result in the unification of features coming from the verb form and the independent pronoun.

For example, assume that an independent subject pronoun occurs after an analytic verb form, as in (10). Then the person and number information of the two words are matched against each other, using these templates.

(10) *Tuigeann mé an fhadhb.
understand.Pres I ART problem
'I understand the problem.'

The analysis of (10) involves the lexicon entry of *mé* 'I' as given in (6), which assigns the value `1` to the feature `PERS`. It also involves the verb form *tuigeann*, which, according to the FST analysis in (4), can be either third person or second person, singular or plural. The unification and hence the parse consequently fail, as the template `AGR-P` tries to assign either third person or second person to the subject, while the lexicon entry for *mé* tries to assign first person to the subject. Figure 1 shows one of the failed parses where XLE tries to unify first person information with second person information.

```
[PRED   'tuig<[17:PRO], [1-OBJ]>'
 SUBJ   [PRED 'PRO'
         PERS {[1]
               [2]}
     17 [CASE com, NUM sg, PRON-TYPE pers]
 OBJ    [CASE com]
 1 [TENSE pres]]
```

Figure 1: Failed parse of the sentence in (10)

If the information coming from the verb form and the subject pronoun matches, the parse succeeds. In (11), the person and number features of the subject pronoun *sé* agree nicely with the person and number features of the analytic verb form.

(11) Tuigeann sé an fhadhb.
understand.Pres he ART problem
'He understands the problem.'

The analysis produced by the computational grammar for (11) is shown in Figure 2.[3]

```
"tuigeann sé an fhadhb."
[PRED   'tuig<[17:PRO], [49:fadhb]>'
 SUBJ   [PRED 'PRO'
     17 [CASE com, NUM sg, PERS 3, PRON-TYPE pers]
 OBJ    [PRED 'fadhb'
         SPEC [DET {[PRED 'an'
                46  [DET-TYPE def, GEND fem]]}]
     49 [CASE com, DEF +, GEND fem, NUM sg, PERS 3]
 1 [CLAUSE-TYPE decl, TENSE pres]]
```

Figure 2: Valid parse of the sentence in (11)

3.3 Evaluation

For evaluation purposes, I manually constructed a testsuite of 30 grammatical and ungrammatical sentences. The implementation currently includes present tense and preterite verb forms in all paradigms and can very easily be extended to include other tenses. The implementation obtains full coverage of the testsuite sentences without any overgeneration.

4 Conclusion

I have presented data from Irish demonstrating the problem of analytic and synthetic verb forms. I have described two earlier approaches; one does not offer an implementation, the other one does offer an implementation, but involves inefficient lexicon checking. I have described my own implementation, which is done using a detailed finite-state morphological analyzer and a computational LFG grammar. The grammar uses efficient templates which rule out non-agreeing verb-pronoun combinations, thereby effectively blocking analytic verb forms where synthetic ones are available.

[3] One reviewer asks about the speed of the implementation. XLE consists of cutting-edge algorithms for parsing and generation using LFG grammars. It is the basis for the ParGram project, which is developing industrial-strength grammars for a variety of languages. XLE returns the following figures after parsing the sentence in (11):

```
1 solutions, 0.020 CPU seconds, 0.000MB max mem, 42
subtrees unified
```

The sentence has 1 solution, it took XLE 0.020 CPU seconds to parse it, and 42 subtrees were unified during the parse.

References

Avery D. Andrews. 1990. Unification and Morphological Blocking. *Natural Language and Linguistic Theory*, 8:507–557.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Publishers.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Miriam Butt. 2007. The Role of Pronominal Suffixes in Punjabi. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, editors, *Architectures, Rules, and Preferences*. CSLI Publications.

Dick Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman, 2008. *XLE Documentation*. Palo Alto Research Center.

Julie Anne Legate. 1999. The Morphosyntax of Irish Agreement. *MIT Working Papers in Linguistics*, 33.

James McCloskey and Kenneth Hale. 1984. On the Syntax of Person-Number Inflection in Modern Irish. *Natural Language and Linguistic Theory*, 1(4):487–534.

Ingo Mittendorf and Louisa Sadler. 2006. A Treatment of Welsh Initial Mutation. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG06 Conference*.