On the Syntax and Translation of Finnish Discourse Clitics*

Aarne Ranta

University of Gothenburg

Lauri Carlsonille.

Motto

> *One criterion is to think of the description as material for machine translation—that is the level of specificity I'd like to achieve. The description of the clitics should support translation between correct uses of clitics and corresponding devices in other languages.* (Carlson 1993, p. 5)

1 Introduction

Finnish has a set of morphemes called **discourse clitics**, attached to words in a way typical of clitics (Zwicky 1977). Some of these clitics attach to the first constituent of a clause, to express things like the formation of questions (*ko*, much like the Latin clitic *ne*), contrasting (*pas*), and reminding (*han*).[1] These three functions are illustrated by the following examples:

> *Jussi juo maitoa* (no clitic, neutral): "John drinks milk"
> *Jussiko juo maitoa* (*ko*, question): "is it John who drinks milk"
> *Jussipas juo maitoa* (*pas*, contrasting): "it is John who drinks milk (and not Peter)"
> *Jussihan juo maitoa* (*han*, reminding): "as we know, John drinks milk"

There is yet another clitic, *kin*, which can often be directly translated by "also" or "even". It attaches to (almost) any element in a clause:

> *Jussikin juo maitoa nykyään* (subject) "also John drinks milk nowadays (and not only Peter)"
> *Jussi juokin maitoa nykyään* (verb): "John even drinks milk nowadays (and not only produces it)"

* Draft of a paper appeared in D. Santos, K. Lindén and W. Ng'ang'a (eds), *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday.* Springer, Heidelberg, 2012. pp. 227–241.

[1] The full list is *ko*, *pa*, *han*, and *s*. Also the combinations *kos*, *kohan*, *pas*, and *pahan* exist. The ones we study here are chosen because of their high frequency and clearly distinguishable meanings.

Jussi juo maitoakin nykyään (object): "John drinks also milk nowadays (and not only beer)"
Jussi juo maitoa nykyäänkin (adverb): "John drinks milk nowadays too (and not only in the past)"

In this paper, we shall give a set of formal rules for how the discourse clitics may appear in Finnish sentences: their *syntax*. We shall also take a look at the *translation* of discourse clitics to English, in some of their typical uses. To satisfy the motto of this paper, we have built a system that performs the translation automatically, in both directions. An on-line demo and its source code can be found on the web.[2]

There is a considerable literature around Finnish discourse clitics. We have been particularly inspired by Karttunen and Karttunen (1976), Nevis (1986) and Carlson (1993). Karttunen and Karttunen (1976) concentrate their study on the clitic *kin*, giving a detailed account of its syntax and semantics in a Montague-style grammar (Montague 1974). Carlson (1993) addresses all clitics (but less formally), showing how they can be interpreted and translated in a discourse context following the idea of dialogue games (Carlson 1983). Nevis (1986) is a thorough linguistic study placing the Finnish clitics in the context of a general theory of clitics (Zwicky 1977).

This paper can be seen as a further development of the Montague-style grammar of Karttunen and Karttunen (1976): the rule system is extended from *kin* to the other clitics and integrated in a wide-coverage resource grammar of Finnish (Ranta 2009). The grammar is formalized and implemented by using the grammar formalism GF (Grammatical Framework, Ranta 2004, 2011), which is designed for supporting **multilingual grammars**. The translation system we present is by definition **compositional**, in the sense that the Finnish and English sentences have a rule-to-rule correspondence via a common tree structure, an **abstract syntax**.

Of course, we can only scratch the surface of the translation of discourse clitics in this paper. One reason, repeatedly shown in Carlson (1993), is that the clitics have several functions, and they can only be disambiguated in the context of a dialogue. For instance, *kin* can be used to express "also" (as above and in Karttunen and Karttunen 1976), but it also has the function of expressing surprise. Thus *Jussi juokin maitoa nykyään* has another translation, "John drinks milk nowadays, after all" (following Carlson 1993). Current machine translation methods are just incapable of selecting between these alternatives in an informed way, as they work sentence by sentence.

A quick experiment with standard machine translation systems confirms how far they are from coping with with Finnish discourse clitics. Google Translate[3] often returns Finnish words with clitics untranslated—just because many word+clitic combinations have never appeared in the training corpus. The Sunda system[4] tailored for Finnish does a better job in rendering Finnish clitics in En-

[2] http://www.grammaticalframework.org/demos/finnish-clitics/
[3] `http://translate.google.com`
[4] `http://www.sunda.fi/eng/translator.html`

glish. But when the English sentences are translated back to Finnish, the clitics disappear. Instead, literal translations of the English sentences are returned. For instance, *Jussikin juo maitoa* is correctly translated to *also Jussi drinks milk*, but the back-translation is *myös Jussi juo maitoa*, which is correct but uses the adverb *myös* ("also") instead of the clitic.

This suggests a conjecture that the frequency of discourse clitics can be used for distinguishing native Finnish from "translationese". This may even be a characteristic of foreign speakers' Finnish, even fluent ones'. Discourse clitics (with the exception of the question clitic *ko*) can *always* be avoided by using paraphrases. When a source text, or a foreign speaker's "mentalese", is being rendered into Finnish (or any other language), the translator/speaker performs a search for an adequate rendering of its meaning. This search has a (legitimate) tendency to return the syntactically closest translation variant. Rendering an English dialogue with Finnish discourse clitics requires a translator with the Finnish speaker's intuitions who continuously poses the question, "how would I express this if I was in the same situation". Then the clitics will naturally appear in many cases.

2 The abstract syntax of discourse clitics

We will focus on two groups of discourse clitics: *han* and *pas* on the one hand, and *kin* on the other. The clitics *han* and *pas* are always attached to the first constituent of a clause (Nevis 1986, Carlson 1993). The clitic *kin* can be attached to *any* constituent. Both groups include some other clitics, too, as mentioned in Section 5 below.

One of the facts we need to formalize is that any sentence contains at most one clitic from each group. They appear in positions that we will call the **topic** and the **focus**; these are not always exactly the traditional semantic topic and focus, which also depend on other things such as intonation and further details of word order[5]. Thus for us, the topic is simply the fronted element, and it may carry *han* or *pas*. The focus is simply any element (including the topic itself) that may carry the focus clitic *kin* (sometimes in combination with a topic clitic). For instance, *maitoakinhan Jussi juo*, "as we know, John drinks milk too", has *maitoa* ("milk", partitive case) as both topic and focus.

What are the "elements" of a clause? We will distinguish four elements: the **subject**, the **verb**, the **object**, and the **adverb** (this will be generalized in Section 5). Any of these elements can work as both topic and focus in the way described above.[6]

[5] Also the Finnish reference grammar Hakulinen & al. (2005) calls *kin* a focus particle, whereas the others are called "tonal particles" ("sävypartikkeli").

[6] The verb doesn't easily get the focus clitic when topicalized: *juokin Jussi maitoa* ("Jussi actually does even drink milk") is strange. On the other hand, *taidankin tästä lähteä* ("I think I leave now") is correct, maybe because the subject is omitted. *Tulikin talvi* ("the winter came, after all") is also correct, maybe because there is an omitted formal subject different from *talvi* ("the winter"). We will leave room for overgeneration here to keep the rules simple.

Our grammar has seven syntactic categories, defined as follows in GF:

```
cat
  S ;        -- declarative sentence
  Clause ;   -- clause with focus on some element (or none)
  Elements ; -- clause elements: subject, verb, object, adverb
  Clitic ;   -- discourse clitic: "han", "pas"
  NP ;       -- noun phrase
  V2 ;       -- two-place verb
  Adv ;      -- adverb
```

The keyword `cat` starts a group of **category declarations**. Each declaration above has a comment (started by a double dash) explaining what the category is meant for.

The `cat` declarations belong to the **abstract syntax** of a GF grammar, similar to the level of "analysis trees" in Montague grammar. In addition to the categories, an abstract syntax contains **function declarations** (`fun`), defining how to construct **abstract syntax trees**. The following five `fun` declarations define five ways of building a top-level sentence from a topic clitic and a clause:

```
fun
  NoTop :                    -- Jussi juo maitoa nyt
    Clitic -> Clause -> S ;  -- John drinks milk now
  TopSubj :                  -- Jussi maitoa juo nyt
    Clitic -> Clause -> S ;  -- it is John who drinks milk now
  TopVerb :                  -- juo Jussi maitoa nyt
    Clitic -> Clause -> S ;  -- John actually does drink milk now
  TopObj :                   -- maitoa juo Jussi nyt
    Clitic -> Clause -> S ;  -- milk is drunk by John now
  TopAdv :                   -- nyt Jussi juo maitoa
    Clitic -> Clause -> S ;  -- now John drinks milk
```

There is thus one rule for topicalizing each of the elements of a clause, plus a "neutral" rule. Since both the neutral rule and the subject topicalization front the subject, we distinguish the latter by moving the verb after the subject; this seems to capture well the idea of topicalizing the subject.[7]

Clauses are formed in two steps. The **predication** step collects the elements together and chooses their proper forms, in terms of agreement. The **focusing** step brings one of the parts of a clause into focus. It can also say that there is no focus (i.e. no *kin*).

```
fun
  Pred : NP -> V2 -> NP -> Adv -> Elements ;
```

[7] Many other permutations are possible, since Finnish has "free word order". Notice, however, that this does *not* mean free variation, since each word order has its own meaning and may, consequently, have its own translation.

```
  NoFoc :                -- Jussi juo maitoa nyt
    Elements -> Clause ; -- John drinks milk now
  FocSubj :              -- Jussikin juo maitoa nyt
    Elements -> Clause ; -- even John drinks milk now
  FocVerb :              -- Jussi juokin maitoa nyt
    Elements -> Clause ; -- John even drinks milk now
  FocObj :               -- Jussi juo maitoakin nyt
    Elements -> Clause ; -- John drinks milk too now
  FocAdv :               -- Jussi juo maitoa nytkin
    Elements -> Clause ; -- John drinks milk now too
```

As we are only using *kin* as the focus clitic, we don't have an argument place for it.[8] But we do need to define the topic clitics, including the absence of one:

```
  noClitic       : Clitic ;  -- (empty)
  remindClitic   : Clitic ;  -- han / as we know
  contrastClitic : Clitic ;  -- pas / no (but)
```

Finally, to test the grammar with actual examples, we define a small lexicon:

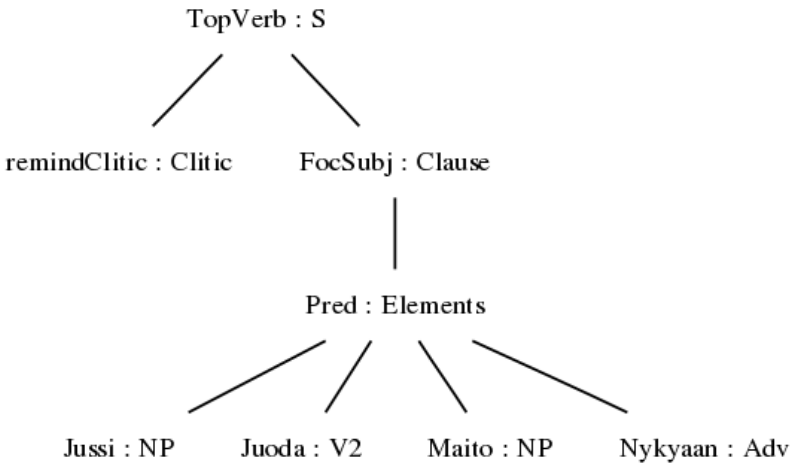
```
fun
  Jussi, Marja, Maito, Viini : NP ; -- John, Mary, milk, wine
  Juoda, Rakastaa : V2 ;            -- drink, love
  Nykyaan : Adv ;                   -- nowadays
```

The abstract syntax we have defined allows us to build 2,400 abstract syntax trees (75 sentence forms times 32 combinations of elements). One example is

```
  TopVerb remindClitic (FocSubj (Pred Jussi Juoda Maito Nykyaan))
```

The tree visualization tool of GF can show it in a nicer form:

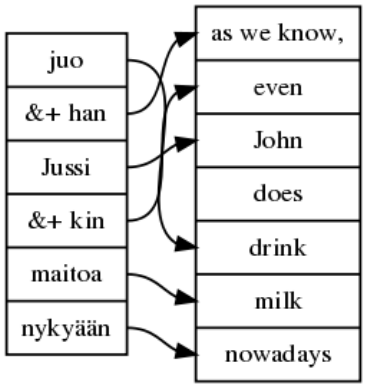
[8] Adding the other clitic of this class, *kaan*, will not change this, since it is in complementary distribution with *kin* depending on the polarity of the sentence; the positive *kin* is "also", and the negative *kaan* is "either". Negative polarity is usually expressed by sentence negation, but can also appear in unnegated questions.



This tree corresponds to the Finnish and English sentences

juohan Jussikin maitoa nykyään
as we know, even John does drink milk nowadays

These translations are produced by **concrete syntaxes** of the abstract syntax. A concrete syntax is a compositional, reversible mapping from trees into strings (and other structures) of a language. The abstract and concrete syntaxes together define a relation of **phrase alignment** between the translations. For the example at hand, the visualization tool of GF gives the following result:



In the picture, "&+" is the **binding operator** that glues the clitic to the foregoing word (see next Section).

3 The Finnish concrete syntax

A concrete syntax defines, for each language separately, how the trees of an abstract syntax are **linearized**. The output of linearization is often a string,

but it can also be a richer data structure. GF has two such structures: **tables** and **records**.

A table is like an *inflection table* in traditional grammar: it gives values to every element in a finite **parameter set**. In the current fragment of Finnish, we use two parameter sets, defined as follows in GF:

```
param
  Case = Nom | Part ;        -- case: nominative or partitive
  Harmony = Back | Front ;   -- vowel harmony: back or front
```

An example of a table is the clitic *han*, which has the form *hän* when attached to a word requiring front vowel harmony. We can define this clitic as a constant whose type is a table type,

```
  han : Harmony => Str = table {Back => "han" ; Front => "hän"}
```

Similarly, noun phrases are tables depending on case,

```
  maito : Case => Str = table {Nom => "maito" ; Part => "maitoa"}
```

So, how do we combine a noun and a clitic? We take the noun in any of the case forms and then attach a clitic, whose form depends on the harmony of the noun: *maitoa+han* ("milk as we know") but *viiniä+hän* ("wine as we know"). A way to implement this in GF is to equip each word in the lexicon with information about its vowel harmony. We do this by means of the record type "string with harmony", defined as

```
  oper HStr : Type = {s : Str ; h : Harmony} ;
```

(where `oper` stands for auxiliary operations). When we have a string with a harmony, we can combine it with a harmony-dependent table by using the following operation:

```
  oper harmony : HStr -> (Harmony => Str) -> Str =
    \hs,ht -> hs.s ++ ht ! hs.h ;
```

In words: we concatenate (`++`) the `s`-field `hs.s` of the harmony-providing string `hs` where we select (`!`) the `h`-field from the harmony-dependent table `ht`.

To make the harmony explicit for noun phrases, we change their type from `Case => Str` to `Case => HStr`. Thus the harmony of a noun depends also on its case. In practice, it is almost always the same for all cases for a given noun, but there are exceptions such as *meri*(+*hän*), *merta*(+*han*) ("see", nominative and partitive).[9]

[9] There are two other ways of dealing with the vowel harmony of clitics in GF. One is to introduce the clitics as forms in inflection tables directly. This, however, leads to prohibitively large tables—for instance, every noun then has almost 3,744 forms (26 case-number combinations, 6 possessive suffixes (incl. none), 3 focus clitics (*kin*,

The parameter types and data structures are used for defining **linearization types** for each category in the abstract syntax. As the linearization types belong to the concrete syntax, they are language-dependent. Here are the linearization type definitions (`lincat`) in Finnish:

```
lincat
  S = Str ;
  Clause, Elements = {subj,verb,obj,adv : HStr} ;
  Clitic = Harmony => Str ;
  NP  = Case => HStr ;
  V2  = HStr ;
  Adv = HStr ;
```

Thus sentences are linearized to plain strings (`Str`). Clauses are records with separate strings for each of the four components. Clitics are tables depending on vowel harmony. The rest of the categories are strings with harmony, which is needed when combining them with clitics. Noun phrases moreover depend on case.[10]

For each function (`fun`) in the abstract syntax, the concrete syntax gives a **linearization rule** (`lin`). Here are the rules for the sentence-forming functions.

```
lin
  NoTop pa c =
    harmony (c.subj) pa ++ c.verb.s ++ c.obj.s  ++ c.adv.s ;
  TopSubj pa c =
    harmony (c.subj) pa ++ c.obj.s  ++ c.verb.s ++ c.adv.s ;
  TopVerb pa c =
    harmony (c.verb) pa ++ c.subj.s ++ c.obj.s  ++ c.adv.s ;
  TopObj pa c =
    harmony (c.obj)  pa ++ c.verb.s ++ c.subj.s ++ c.adv.s ;
  TopAdv pa c =
    harmony (c.adv)  pa ++ c.subj.s ++ c.verb.s ++ c.obj.s ;
```

kaan, none), and 8 topic clitics (all combinations incl. none)); the number of distinct forms is a little lower, since some of the combinations of case and possessive suffix produce the same string. The other way is to leave the decision to a separate lexical synthesis procedure (unlexing) after grammar-based linearization. This helps keep the grammar simple, but makes the over-all system more complex. One complication is that the vowel harmony of compound nouns, which are very common in Finnish, is impossible to decide from a string alone, without knowing the compound boundary. The parameter-based all-GF solution used here gives good quality with a reasonable table size. The classic implementation of Finnish morphology by Koskenniemi (1983) treats clitics as lexical forms to preserve accuracy, but avoids the explosion of the lexicon because its run-time representation is a finite-state automaton rather than an explicit table. Our solution similarly results in an automaton at run time, if we add a lexical analysis phase needed for restoring the binding tokens following the ideas of Huet (2005).

[10] A full Finnish grammar has many more dependencies, in particular for verbs; even nouns have 30 forms in the GF resource grammar.

Each rule expresses topicalization by fronting one of the elements. This element is combined with the topic clitic by using the `harmony` function to select the proper form of the clitic.

The predication rule selects the proper forms of the constituents; here we only need to select the case of the subject and (partitive) object:

```
lin Pred subj verb obj adv =
  {subj = subj ! Nom ; verb = verb ; obj = obj ! Part ; adv = adv} ;
```

The focus rules put the focus clitic *kin* in place if needed. They use the auxiliary operation `kin`, which attaches the *kin* clitic to an `HStr`:

```
oper
  kin : HStr -> HStr = \hs -> {s = hs.s ++ bind "kin" ; h = hs.h} ;
lin
  NoFoc c = c ;
  FocSubj c =
    {subj = kin c.subj ; verb = c.verb ; obj = c.obj ; adv = c.adv} ;
  FocVerb c =
    {subj = c.subj ; verb = kin c.verb ; obj = c.obj ; adv = c.adv} ;
  FocObj c =
    {subj = c.subj ; verb = c.verb ; obj = kin c.obj ; adv = c.adv} ;
  FocAdv c =
    {subj = c.subj ; verb = c.verb ; obj = c.obj ; adv = kin c.adv} ;
```

It remains to linearize the clitics and the test lexicon. For the clitics, we define an auxiliary similar to `mkClause`:

```
  oper mkClitic : Str -> Str -> Harmony => Str =
    \ko,koe -> table {Back => ko ; Front => koe} ;
```

Using this, we define

```
lin
  noClitic        = mkClitic [] [] ;
  remindClitic    = mkClitic (bind "han") (bind "hän") ;
  contrastClitic  = mkClitic (bind "pas") (bind "päs") ;
```

The `bind` operation adds the binding token `&+`,

```
  oper bind : Str -> Str = \s -> "&+" ++ s ;
```

The binding token is eliminated by an **unlexer**, a post-processing phase after linearization. It produces

```
  Jussi &+ han juo maitoa &+ kin --> Jussihan juo maitoakin
```

As a preprocessing phase before the parser, the **lexer** recognizes possible clitics and introduces binding tokens,

```
Jussikinhan juo maitoa --> Jussi &+ kin &+ han juo maitoa
```

When defining the lexicon, we don't want to give the vowel harmony of each word explicitly, but infer it with a simple heuristics, which inspects a string and determines it as a back vowel string if and only if it includes *a*, *o*, or *u*. This operation is definable in GF by regular-expression pattern matching:

```
oper mkHStr : Str -> HStr = \s -> {
       s = s ;
       h = case s of {
         _ + ("a" | "o" | "u") + _ => Back ;
         _ => Front
         }
       } ;
```

Since verbs and adverbs are plain `HStr`'s, just `mkHStr` is needed to define them compactly in the lexicon. For noun phrases, we use a derived operation,

```
oper mkNP : Str -> Str -> Case => HStr =
  \n,p -> table {Nom => mkHStr n ; Part => mkHStr p} ;
```

Now we can define the lexicon compactly:[11]

```
lin
  Jussi    = mkNP "Jussi" "Jussia" ;
  Maito    = mkNP "maito" "maitoa" ;
  Marja    = mkNP "Marja" "Marjaa" ;
  Viini    = mkNP "viini" "viiniä" ;
  Juoda    = mkHStr "juo" ;
  Rakastaa = mkHStr "rakastaa" ;
  Nykyaan  = mkHStr "nykyään" ;
```

4 The English concrete syntax

The abstract syntax in Section 2 was designed to account for Finnish discourse clitics. Can we map it into English in a compositional way? This turned out to be easy, even though the result is somewhat arbitrary: sure there can be other English translations, some equivalent and some corresponding to different interpretations of the clitics. But the translations chosen here suggest that any other ones could be defined in similar, compositional ways.

Let us assume that the abstract syntax in Section 2 encodes a fixed set of meanings—in particular, that the function `remindClitic` is used for reminding

[11] The Finnish resource grammar uses regular-expression pattern matching to define a set of much more powerful lexical paradigms, which infer the complete inflection from just the dictionary form for 87% of nouns and 96% of verbs (Détrez and Ranta 2012).

and `contrastClitic` for contrasting, and that the focus clitic *kin* means "too", or "even". We will give just one English translation to each construction, aimed to be among the much larger set of semantically faithful and stylistically correct translations. Carlson (1993) uses many more variants to achieve a livelier style.

As English has less inflection than Finnish and no vowel harmony, some linearization types are simpler. But in English we need variation in verb forms. We say *John drinks milk* (third person singular present indicative) in normal cases, *John does drink milk* (infinitive) to topicalize the verb, and *milk is drunk by John* (past participle) to topicalize the object. Since the form can only be selected on the sentence (`S`) level, clauses must use verb inflection tables rather than plain strings.

```
param
  VForm = Inf | Ind | PPt ;
lincat
  S = Str ;
  Clause, Elements = {subj,obj,adv : Str ; verb : VForm => Str} ;
  Clitic = Str ;
  NP  = Str ;
  V2  = VForm => Str ;
  Adv = Str ;
```

As the "corresponding devices" to Finnish topicalization, we will use *it* clefts for the subject, the auxiliary *do* for the verb, passive voice for the object, and plain fronting for the adverb. Reminder is expressed by *as we know* and contrast by a leading *no*. Here are the sentence-forming rules:

```
lin
  NoTop pa c =
    pa ++ c.subj ++ c.verb ! Ind ++ c.obj ++ c.adv ;
  TopSubj pa c =
    pa ++ "it is" ++ c.subj ++ "that" ++ c.verb ! Ind ++ c.obj ++ c.adv ;
  TopVerb pa c =
    pa ++ c.subj ++ "does" ++ c.verb ! Inf ++ c.obj ++ c.adv ;
  TopObj pa c =
    pa ++ c.obj  ++ "is" ++ c.verb ! PPt ++ "by" ++ c.subj ++ c.adv ;
  TopAdv pa c =
    pa ++ c.adv  ++ c.subj ++ c.verb ! Ind ++ c.obj ;

  noClitic        = [] ;
  remindClitic    = "as we know," ;
  contrastClitic  = "no," ;
```

In the clause-forming rules, we use *even* to translate *kin* for the "earlier" elements (subject and verb), and *too* for the "later" ones (object and adverb). This gives a good approximation of what sounds natural.

```
lin
  Pred subj verb obj adv =
    {subj = subj ; verb = verb ; obj = obj ; adv = adv} ;

  NoFoc c = c ;
  FocSubj c = {subj = "even" ++ c.subj ; verb = c.verb ;
               obj = c.obj ; adv = c.adv} ;
  FocVerb c = {subj = c.subj ; verb = \\f => "even" ++ c.verb ! f ;
               obj = c.obj ; adv = c.adv} ;
  FocObj  c = {subj = c.subj ; verb = c.verb ;
               obj = c.obj ++ "too" ; adv = c.adv} ;
  FocAdv  c = {subj = c.subj ; verb = c.verb ; obj = c.obj ;
               adv = c.adv ++ "too"} ;
```

The lexicon is simple to define:

```
lin
  Jussi    = "John" ;
  Maito    = "milk" ;
  Marja    = "Mary" ;
  Viini    = "wine" ;
  Juoda    = mkVerb "drink" "drunk" ;
  Rakastaa = mkVerb "love" "loved" ;
  Nykyaan  = "nowadays" ;
oper
  mkVerb : Str -> Str -> VForm => Str =
    \s,p -> table {Inf => s ; Ind => s + "s" ; PPt => p} ;
```

5 Scaling up

We have given the complete source code of a toy grammar that translates between Finnish sentences with discourse clitics and English sentences with corresponding devices. Choosing to work on a toy grammar has made it possible to give the complete details, and also to focus on the critical issues.

The main issue we have addressed is the combinatorics of the discourse clitics, dealt with by the use of the clause records {`subj,verb,obj,adv : Str`} in both Finnish and English (with slight variations). The whole account relies on the use of a record data structure, rather than a plain string, as the target of linearization. The elements of the record can then be focalized, topicalized, and otherwise reordered in different ways. This structure is inspired by the **topological structure** of Germanic languages (Diderichsen 1962). The rationale is the same in Finnish as in German and Danish: the use of **discontinuous constituents** exemplified by the topological structure makes it possible to reorder the parts of a clause to express discourse structures.[12]

[12] In a wide perspective, our approach can be seen in relation to the "quantifying in" idea of Montague (1974), which was developed for the clitic *kin* in Karttunen

Another, minor, issue is the treatment of inflection and vowel harmony in Finnish. We have wanted to show how a lexicon can be efficiently built by an underlying morphological machinery and high-level functions that hide it from the user (here, `mkNP` and `mkHStr`). We have also shown that the choice of the correct form of a clitic can be performed accurately by memorizing the vowel harmony of each word.

While being demonstrated in a toy grammar, the approach used here is very much the same as in the full-scale resource grammar for Finnish. The **core resource grammar** (Ranta 2009) is an implementation of an abstract syntax of around 80 categories, 120 combination rules, and 500 lexemes. The core grammar is completed by a GF version of the KOTUS word list of 77,000 lexemes.[13] The core resource grammar implements a set of syntactic structures for 22 languages (in February 2012). Finnish was one of the first languages implemented (starting in 2003), and certainly did have some influence on the design of the abstract syntax. However, the core abstract syntax encodes a kind of "Standard Average European" and doesn't, in particular, cover the discourse clitics so peculiar to Finnish.[14]

In contrast to the toy grammar, the resource grammar version of clitics aims to cover their syntax completely. Thus it adds, among other things,

- the full set of topic clitics and their combinations (adding *pa*, *pahan*, *ko*, *kos*, *kohan*);
- the focus clitic *kaan* and its complementary distribution with *kin* triggered by negative polarity: *Jussi juo maitoakin* ("John drinks milk too") vs. *Jussi ei juo maitoakaan* ("John doesn't drink milk either");
- the interplay with negation and tenses, including the fronting of the negation (*eihän Jussi juo maitoakaan* "as we know, John actually doesn't drink milk either");
- other forms of clauses than just subject-verb-object-clitic;
- larger lexicon, with the generalizations it requires in syntax (e.g. the infamous Finnish object case, which persists in discourse rearrangements).

As an example of what these extensions involve, let us look at the linearization type of clauses:

```
{s : Tense => Polarity => {subj,fin,inf,obj,adv,ext : HStr}}
```

This record generalizes our toy grammar in two ways. First, it has six fields instead of four: the `verb` field is split into a finite and an infinite part (`fin`,

and Karttunen 1976. The common idea is that the clitic doesn't primarily attach to a word, but to an entire clause, from which a selected word is picked for the final, concrete attachment. Rather than bound variables, we use the idea of "slash categories" of GPSG (Gazdar & al. 1985): categories that have "gaps" in which syntactic constructions can insert new material.

[13] `http://kaino.kotus.fi/sanat/nykysuomi/`

[14] Other "non-standard" languages represented in the resource grammar library are Amharic, Arabic, Hindi/Urdu, Maltese, Nepali, Persian, Punjabi, Swahili, and Thai.

`inf`), and an `ext` field is added for extensions, such as subordinate clauses and extra adverbs. The extensions are never considered for focus or topic. Thus, when building a clause, there is a choice whether an adverb (or a complement) is placed in the `adv` or the `ext` field. The placement of clitics within constituents can also be controlled at the construction phase, e.g. *tuoretta+kin maitoa* vs. *tuoretta maitoa+kin* (roughly, "even fresh milk (and not only sour milk)" vs. "fresh milk too (and not only fresh bread)"). What is needed is simply that the constituents from which clauses are built, such as noun phrases, are themselves discontinuous, and are stored as records rather than strings in the clause.

The second generalization is that tense and polarity may be varied. When full sentences are built from clauses, any of the first five fields can be focalized and topicalized, with some restrictions depending on tense and polarity. For instance, the negation verb *ei* (which work's much like *don't* in English), cannot be focalized.[15]

At the time of writing, the resource grammar version of discourse clitics does not yet cover other than declarative sentences, except questions with the standard question clitic *ko*. Some of the sub-clausal discontinuities are not yet covered either. The lexical treatment of vowel harmony is not carried out in all parts of speech; adapting the treatment used here would also help with other clitics, such as the possessive suffix *nsa-nsä* ("his", "her").

The resource grammar of GF is *not* meant to be used as an interlingua for translation, but as a library for implementing concrete syntaxes of more restricted and fine-grained interlinguas. In typical GF applications, an interlingua is specific to a domain, which can range from mathematical proofs to touristic phrases. The interlingua presented in this paper, however, is not domain-specific. When equipped with a large lexicon, it could therefore be able to produce good translations of a large set of sentences whose structures are among the ones treated here.

6 Conclusion

We have discussed the combinatorics of Finnish discourse clitics and shown in full detail a toy grammar formalizing them for a fragment of language. The grammar was given an abstract syntax that permits compositional translation to other languages, which was illustrated by English. The English grammar uses different means (adverbs, fronting, passives, *it* clefts, emphatic *do*) to express the same things as Finnish expresses by discourse clitics. We have also summarized the main issues in the generalization of the toy grammar into a component of a wide-coverage Finnish resource grammar.

[15] Carlson 1993 presents this as a consequence of the general rule that "*-kin/-kAAn* cannot modify the polarity alone". Interestingly, this rule seems to be getting less strict, at least for two-syllabic plural forms: Google search finds e.g. the natural-sounding *Maapallo kyllä selviää, vaikka me emmekään selviäisi* ("The globe will certainly survive, even if we didn't survive ourselves", web version of the newspaper *Keskisuomalainen*, May 2008).

The toy grammar is intended to serve as a prototype of a system able to translate "between correct uses of clitics and corresponding devices in other languages", a goal stated in Carlson (1993). While the resource grammar version of the system already covers a wide range of syntactic combinations, we have not formalized the semantic distinctions between different uses of the clitics. Thus we haven't addressed the disambiguation problem, which is a task that seems to need a wider context than the isolated sentence—a dialogue game, as suggested by Carlson.

We don't claim to have solved deep linguistic problems or even taken into account all the theoretical findings that have been made about Finnish discourse clitics. But we have shown one way in which the clitics can be painlessly integrated in formal syntax and lead to running implementations of translation systems. The demo system accompanying this paper seems to be the first one that translates the clitics from Finnish to English without loss, and even produces them when translating from English to Finnish. While it is a system of a minuscule scope, it can be useful for tasks such as language training for learners of Finnish. It can for instance be used in the quiz mode, where the user sees an English sentence and is invited to construct a Finnish translation.[16]

Acknowledgements

I am grateful to Janet Pierrehumbert and Atro Voutilainen for useful and encouraging comments on the first version of this paper.

[16] See `http://www.grammaticalframework.org/demos/finnish-clitics/`

Bibliography

Carlson, L. (1983). *Dialogue Games: An Approach to Discourse Analysis*. Dordrecht: D. Reidel Co.

Carlson, L. (1993). Dialogue Games with Finnish Clitics. In M. Vilkuna and S. Shore (Eds.), *Yearbook of the Linguistic Society of Finland*. Helsinki: SKY.

Détrez, G. and A. Ranta (2012). Smart paradigms and the predictability and complexity of inflectional morphology. In *EACL 2012*.

Diderichsen, P. (1962). *Elementaer dansk grammatik*. Kobenhavn.

Gazdar, G., E. Klein, G. Pullum, and I. Sag (1985). *Generalized Phrase Structure Grammar*. Oxford: Basil Blackwell.

Hakulinen, A., M. Vilkuna, R. Korhonen, V. Koivisto, T. R. Heinonen, and I. Alho (2005). *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Huet, G. (2005). A Functional Toolkit for Morphological and Phonological Processing, Application to a Sanskrit Tagger. *The Journal of Functional Programming 15(4)*, 573–614.

Karttunen, F. and L. Karttunen (1976). The clitic -kin/-kaan in Finnish. *Texas Linguistic Forum 5*, 89–118.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph. D. thesis, University of Helsinki.

Montague, R. (1974). *Formal Philosophy*. New Haven: Yale University Press. Collected papers edited by Richmond Thomason.

Nevis, J. A. (1986). *Finnish Particle Clitics and General Clitic Theory*. Ph. D. thesis, Department of Linguistics, Ohio State University, Columbus.

Ranta, A. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism. *The Journal of Functional Programming 14(2)*, 145–189. http://www.cse.chalmers.se/~aarne/articles/gf-jfp.pdf.

Ranta, A. (2009). The GF Resource Grammar Library. *Linguistics in Language Technology 2*. http://elanguage.net/journals/index.php/lilt/article/viewFile/214/158.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. Stanford: CSLI Publications. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Zwicky, A. (1976). On clitics. *Indiana University Linguistic Club 5*, 89–118.