Referent tracking and pronoun resolution in Finnish

Elsi Kaiser
Department of Linguistics
University of Pennsylvania
ekaiser@ling.upenn.edu

Abstract

This paper presents a referent-tracking and pronoun resolution system for Finnish, a free-word-order, articleless language that poses a challenge for algorithms designed for languages like English that have definite/indefinite articles. To track referents and interpret pronouns in Finnish, this algorithm uses the pragmatically-motivated word order tendencies of Finnish to create an ordered register of pegs (where each peg is associated with an entity in the discourse), ranked according to salience. Pronouns are interpreted as referring to the topmost peg in the register. The algorithm aims to extend and adapt notions from Dynamic Semantics (Groenendijk et al. 1996) and Centering Theory (Grosz et al. 1995) to a typologically different language.

1 Introduction

In this paper,[1] I present an entity-tracking and pronoun interpretation system which extends and elaborates some of the methods of Centering Theory (e.g. Grosz et al. (1995)) and Dynamic Semantics (e.g. Groenendijk et al. (1996)). In order to track referents and interpret pronouns in Finnish, a free word order language without articles, this algorithm uses the pragmatically-motivated word order tendencies of Finnish to create an ordered register of pegs (where each peg is associated with an entity in the discourse), ranked according to salience. In this algorithm, pronouns are interpreted as referring to the topmost (most salient) peg in the register.

The structure of this paper is as follows. First, Section 2 is a review of some previous work, and in Section 3, I present the basics of the algorithm. I discuss why languages like Finnish pose interesting problems for entity-tracking systems in Section 4. Section

[1] Thanks are due to Robin Clark for many comments and ideas. Thanks also to Ellen Prince, Misha Becker, Tonia Bleam, Na-Rae Han, Alexander Koller, Eleni Miltsakaki, Kimiko Nakanishi, Jennifer Venditti and audiences at Penn, Amsterdam and Helsinki. I would also like to thank three anonymous ESSLLI reviewers and three anonymous WEB-SLS reviewers for very useful comments.

5 addresses the details of the current version of the algorithm. Section 6 discusses the distinction between main and subordinate clauses, and Section 7 contains the conclusion, as well as directions for future work. The algorithm presented here is still work in progress, and is best viewed as a first step rather than the final answer.

2 A look at some previous work

The algorithm presented in this paper makes crucial use of the notion of 'pegs' (Groenendijk et al. 1996), which can be thought of as "addresses in memory" and are used to keep track of entities in a discourse. In this model, as in dynamic semantics, when a new (previously unmentioned) entity enters the discourse model for the first time, it is mapped onto a new peg. Pronouns refer to entities that are already linked to pegs.[2] More specifically, in this algorithm, a pronoun is interpreted as an instruction to find the entity linked to the highest-ranked featurally-compatible peg (The method of ranking the pegs is discussed below in Sections 3 and 5).

The notion that a pronoun 'points to' the topmost peg of an ordered register of pegs is based on Centering Theory (Grosz et al. 1995).[3] According to Centering Theory, which is a model of the local-level component of attentional state in discourse (Grosz et al. (1995:4-5)), the entities ('centers') evoked by a sentence are ordered in terms of their discourse salience. The most salient entity is the one that is considered to be at the center of attention at that particular point in the discourse. The ranking can depend on a number of criteria, including syntactic, morphological and prosodic factors. It is usually assumed to be *subject* > *object* > *others* for English (Walker et al. (1998:7)). Each utterance thus has a single highest-ranked center. In addition, most utterances have a 'backward-looking center', i.e. a center that refers to an entity mentioned in the preceding utterance. According to Centering Theory, if any of the centers expressed in a sentence is a pronoun, then the backward-looking center must be a pronoun ('Pronoun Rule').

3 Introduction to the algorithm

The algorithm, in its current incarnation, is designed to track the entities mentioned in a discourse and to resolve the pronouns that refer to them. It has three main components; (1) A supply of pegs; (2) a discourse register (an array) to act as a storehouse for the pegs that are 'in use' in the discourse, and (3) a current register (an array) for pegs in the current sentence (cf. Hintikka and Sandu (1997)). The pegs are moved from the current

[2] As one of the ESSLLI reviewers noted, this is somewhat of a simplification, since there are some exceptions. For example, after talking about 'a couple', one can then use the pronoun 'she' to refer to the woman in the couple.

[3] See Walker et al. (1998) for a more in-depth look at Centering.

register to the top of the discourse register at the end of every utterance, and their order relative to one another is retained. The current version of the algorithm is not intended to deal with (discourse) deixis. Moreover, at the current stage, only sentences with neutral intonation are considered.[4]

3.1 Ranked pegs

A central aspect of this algorithm is the idea that the pegs are ranked in the registers, with the topmost entity in each register being more accessible[5] than the ones lower down. When a new discourse entity is introduced, it is associated with a peg, and that peg is added to the current register. The algorithm treats subjects and objects asymmetrically to capture the fact that subjects tend to be more salient than objects (as noted for English by Grosz et al. (1995), Hudson-D'Zmura and Tanenhaus (1998), *inter alia*; for crosslinguistic work, see Walker et al. (1998)). To reflect this, the peg for the subject of a sentence is usually located above (i.e. is more accessible than) the peg for the object. This ranking could easily be extended to include other grammatical functions beyond subject and (direct) object. However, for reasons of brevity, this paper only discusses subjects and objects.

It is worth noting that the current version of this algorithm ranks entities in terms of their grammatical functions, not their linear position in the sentence. Using grammatical relations instead of word order as the primary means of ranking referents is supported by some preliminary psycholinguistic results for Finnish (see Kaiser (to appear)), as well as crosslinguistic research in other languages with flexible word order (see e.g. Hoffman (1998), Turan (1998) on Turkish, Prasad and Strube (2000) on Hindi). However, there seems to be some crosslinguistic variation in what factors determine the ranking (see e.g. Rambow (1993), Strube and Hahn (1996), Strube (1998) on German). The algorithm can easily be modified to deal with such differences; for example, it could rank the pegs by linear order or by discourse status (see e.g. Strube and Hahn (1996), Strube (1998)), instead of grammatical function. There is nothing inherent in the algorithm that requires the ranking to be done according to grammatical role.

3.2 Resolving pronouns

According to this algorithm, a pronoun is interpreted as an instruction to point to the topmost peg in the discourse register (actually, to the entity that the topmost peg is connected to) and to bring it into the current register. If the entity linked to the topmost

[4] Note also that this algorithm is not intended to replace Binding Theory. In fact, as one of the reviewers points out, if the input for the algorithm is the output of a parser, then we can assume that coindexation restrictions have already been imposed in the course of the parsing process.

[5] The term 'accessible' is used here to mean 'salient,' where the most salient entity is the one at the center of attention. I chose the term 'accessible' because it captures the notion of pegs being more or less easily acccessed by the algorithm.

peg is featurally incompatible with the pronoun (e.g. the pronoun is *he* but the entity is feminine), then the algorithm checks the next highest peg in the discourse register.

This algorithm is more 'global' than Centering Theory, which focuses on the local relationship between two adjacent utterances within a discourse segment. Within Centering Theory, it is not clear how one should deal with a pronominal 'backward-looking center' that does not have an antecedent in the immediately preceding utterance (e.g. *Peter visited his mother yesterday. She had recently moved to New York. He enjoyed his trip, especially since the weather was so nice.*). In contrast, in the current algorithm, the search for the suitable peg is not limited to the pegs for entities in the immediately preceding utterance. Along similar lines, Walker (1998) argues in favor of a Cache Model, according to which centers can remain accessible even two to three sentences after they occur.

3.3 When to add pegs, when to look for existing pegs

The algorithm needs to ensure that the number of pegs in use is correct. It needs to know whether a given NP, say *cat*, refers to a cat that was already mentioned and has a peg in the discourse register, or whether it is introducing a new discourse entity. In English, the articles *a* and *the* provide useful clues for this task. In contrast, Finnish has neither a definite nor an indefinite article, and thus the distinction between a noun phrase that is 'pegless' and one that already has a peg cannot be made simply on the basis of morphology. Thus, it is not clear how a system, e.g. Dynamic Semantics, that aims to track the referents mentioned in a discourse would deal with Finnish. The algorithm discussed here deals with this complication by using information from word order patterns. The next section addresses the relationship between word order and information status in Finnish.

4 Basics of Finnish word order

As mentioned above, the articleless nature of Finnish poses problems for referent-tracking systems that use the definite/indefinite article distinction to discriminate between expressions referring to new entities and previously mentioned ones.[6] However, the lack of articles does not mean that Finnish fails to mark the distinction between old and new information. Finnish word order is flexible, and can be used to make many of the distinctions that other languages accomplish by using articles (see e.g. Chesterman (1991))[7].

[6] Kruijff-Korbayová (1997) discusses the application of File Change Semantics to Czech, another articleless language, and shows how word order marks the novelty/familarity distinction. However, she does not propose a reference resolution algorithm. Thanks to one of the reviewers for bringing this paper to my attention.

[7] A kind of optional 'definite article ' is evolving in colloquial Finnish from the demonstrative pronoun *se* 'it' (Laury 1997). However, this phenomenon is not a part of standard Finnish.

Although Finnish is canonically SVO, all six permutations of these elements are grammatical in the appropriate contexts (Vilkuna (1995:245)). Different word orders realize different pragmatic structurings of the conveyed information. For example, in orders with more than one preverbal argument (SOV, OSV), as well as verb-initial orders (VOS, VSO), the initial constituent is interpreted as contrastive[8] (see e.g. Vallduví and Vilkuna (1998), Vilkuna (1995)). When trying to ascertain the discourse status of an entity, knowing whether it is contrastive is not necessarily informative because, as Vallduví and Vilkuna (1998) note, the notions of 'rheme' (roughly speaking, 'new' information) and 'kontrast' are distinct and should not be conflated.[9] Thus, SOV and OSV order unfortunately do not tell us whether the initial, contrastive constituent is new or old to the discourse. However, the immediately preverbal constituent in these orders patterns like the preverbal constituent in SVO and OVS order, which *is* informative about discourse-status, as we will see below.

Before turning to the discourse information conveyed by SVO and OVS orders, it is worth noting that verb-initial orders (VSO, VOS) usually have discourse-old arguments. As illustrated by the example below, a verb-initial sentence often has the function of affirming the speaker's belief the proposal under discussion is true (here, whether Mikko gave Anna flowers) (see Vilkuna (1995) for further discussion).

(1) ANTOI Mikko Annalle kukkia. (VS-order, Vilkuna (1995))
Gave Mikko-NOM Anna-ALL flowers-PART
'(Oh yes), Mikko did give Anna flowers.'

Let us now take a closer look at the pragmatics of SVO and OVS orders in Finnish. A preverbal subject NP as well as a preverbal object NP is usually interpreted as being 'old' information, i.e., as referring to an entity already mentioned in the discourse (see Prince (1992) for a discussion of 'discourse-old') (ex. (2), (3)). (As mentioned above, the immediately preverbal argument in SOV and OSV orders matches this pattern, i.e. is interpreted as old information.) If an SVO sentence occurs at the very beginning of a discourse, however, the preverbal subject can be 'new' information. Postverbal NP subjects are 'new' information (ex. (3)). On the other hand, NP objects in their canonical postverbal position can be new or old (ex. (2)). These patterns are summarized in Chart 1 (see Chesterman (1991) for further discussion).

(2) Mies luki kirjan. (SVO order)
Man-NOM read book-ACC
'A/the man read a/the book.'

(3) Kirjan luki mies. (OVS order)
Book-ACC read man-NOM
'The book, a man read.'

[8] I am putting aside here the finer details of the various ways of definining the term 'contrastive.'

[9] See Kaiser (2000a) for further discussion concerning the functions of OSV order in Finnish and the distinction between discourse status and contrast.

Chart (1)

	subject-old	subject new
object-old	SVO	OVS
object-new	SVO	SVO

A corpus study of word order patterns in Finnish (Hakulinen and Karlsson (1979), based on a corpus of 10 000 sentences) found that the majority of sentences, 87 percent, have either SV(O) or (O)VS order. Less than 4 percent of the sentences in the corpus had SOV or OSV order, and less than 4 percent had VSO or VOS order. Thus, even an algorithm which focuses only on SVO and OVS orders as a means of distinguishing old and new entities will still have fairly good coverage. For reasons of exposition, this paper focuses primarily on SVO and OVS, but the reader should keep in mind that, as discussed above, the claims about discourse-status patterns in SVO and OVS orders can be extended to the immediately preverbal arguments in SOV and OSV orders, and that orders with a sentence-initial accented verb usually have discourse-old postverbal arguments.

In the next section we will see how the entity-tracking algorithm can take these ordering patterns into account, in order to avoid misinterpreting previously-mentioned entities as new, or new entities as already mentioned.

5 Algorithm for Finnish

5.1 Full NP subjects and objects

First, we will consider what happens when the algorithm encounters a subject. As illustrated in the simple example below, when the algorithm comes across a preverbal NP subject (‘man’ in ex. (4)) at the very *beginning* of a discourse, it adds a peg to the top of the current register. (Curly brackets { } are used to denote registers, ‘d.r.’ means discourse register, and ‘c.r.’ stands for current register. The rightmost element in each register is the most salient – i.e. ‘topmost.’) When the sentence ends, the peg corresponding to *man* is moved from the current register to the top of the discourse register. The next sentence begins with the (gender-neutral) pronoun *hän* ‘s/he’, which the algorithm interprets as an instruction to look for the highest-ranked featurally-compatible peg. After this peg is located, it is moved to the current register, and the referent linked to this peg is interpreted as the subject of the predicate ‘smiled’. At the end of the second sentence, all pegs in the current register are dumped into the discourse register.

(4)	Mies	käveli	sisään.	Hän	hymyili.	(discourse-initial)
	Man-NOM	walked	in.	S/he-NOM	smiled.	

‘A man walked in. He smiled.’

$\{\ \dots\}_{d.r.}$ $\{\text{man}\}_{c.r.}$	[peg from first sentence]
$\{\text{man}\}_{d.r.}$ $\{\ \dots\}_{c.r.}$	[peg dumped into d.r. at end of sentence]
$\{\dots\}_{d.r.}$ $\{\text{man}\}_{c.r.}$	[pronoun *hän* 's/he' encountered, points to top peg, which is moved to c.r.]
$\{\text{man}\}_{d.r.}$ $\{\dots\}_{c.r.}$	[peg back to d.r. after second sentence is over]

Recall now that a discourse-internal preverbal NP subject usually refers to an already-mentioned entity. Thus, in such a case, the algorithm searches for the peg that already exists for this entity, and brings it to the top of the current register. If the algorithm cannot find a suitable peg, a 'repair' process is presumably triggered; perhaps, as a last resort, the algorithm repairs the situation by adding a new peg. An utterance requiring such a repair is presumably felt to be infelicitous.[10]

When dealing with a postverbal NP subject (new information), the algorithm adds a new peg to the top of the current register.[11] This is illustrated in ex. (5).

(5)	...Naisen	näki	mies.	Hän	oli	iloinen.
	...woman-ACC	saw	man-NOM.	S/he-NOM	was	happy.

'...The woman, a man saw. He was happy.'

$\{\dots\}_{d.r.}$ $\{\text{woman}\}_{c.r.}$	[peg for sentence-initial object moved from d.r. to c.r.]
$\{\dots\}_{d.r.}$ $\{\text{woman, man}\}_{c.r.}$	[peg for postverbal subject added on top of object peg in c.r.]
$\{\text{woman, man}\}_{d.r.}$ $\{\dots\}_{c.r.}$	[pegs dumped into d.r. at end of first sentence]
$\{\text{woman}\ \dots\}_{d.r.}$ $\{\text{man}\}_{c.r}$	[pronoun *hän* 's/he' encountered, peg that pronoun points to is moved to c.r.]
$\{\text{woman, man}\}_{d.r.}$ $\{\dots\}_{c.r.}$	[pegs back to d.r. after second sentence is over]

This example can also be used to illustrate the algorithm's treatment of preverbal objects. When the algorithm comes across a preverbal NP object (old information), it searches for an existing peg in the discourse register, and brings it to the second-highest slot in the current register. At the end of the first sentence, the pegs for the subject and object are thus ranked such that the subject peg is above the object peg. As mentioned earlier, this version of the algorithm treats the subject peg as more accessible than the object peg, regardless of the word order (SVO or OVS). In other words, the algorithm ranks the pegs by grammatical function, not word order or information status. This arrangement predicts that a pronoun in the subsequent sentence will tend to refer to the subject 'man', and not the object 'woman.' Preliminary results from a sentence-completion study (Kaiser (to appear)) provide some support for this prediction for sentences with two full NP arguments, such as ex. (5).

[10] Even an entity which is introduced into the discourse in this infelicitous way can be referred to with a pronoun, which indicates that it is mapped onto a peg and not just 'ignored' by the algorithm.

[11] Here, the term *postverbal* means 'postverbal in the SVO or OVS configuration', and is not intended to describe the behavior of postverbal subjects and objects in VOS or VSO order. To deal with subjects and objects in VSO and VOS orders, i.e. to appropriately interpret them as discourse-old, the algorithm could keep track of how many arguments occur after the verb, and if it finds two postverbal arguments, to treat both as discourse-old.

Interestingly, the results of another sentence-completion experiment (Kaiser (to appear)) suggest that information status (as encoded in word order and NP form) can modulate the effect of grammatical role. In other words, a pronominal, preverbal object which is clearly discourse-old seems to gain somewhat in salience relative to a discourse-new postverbal full NP subject (see Kaiser (to appear) for details). The influence of discourse status on referent salience could be encoded in the algorithm by a more fine-grained ranking mechanism. For example, instead of simply claiming that pegs are ranked by grammatical function, and *subject* > *object*, we could hypothesize that *discourse-old subject* > *discourse-old object* > *discourse-new subject* > *discourse-new object*, where discourse-status is also reflected in the word order. This kind of ranking would capture the effects of both grammatical role and discourse status. However, the exact ranking of discourse-new subjects and discouse-old objects would have to be determined by additional corpus work and/or psycholinguistic research. If we opt to implement this kind of ranking, the algorithm needs some way of determining what is discourse-old and what is discourse-new. A possible search mechanism is described at the end of this section.

It is worth pointing out at this stage that, for any algorithm which uses grammatical function information to rank entities and to resolve pronouns, parallelism effects pose a challenge.[12] Parallelism states that "a pronoun with two or more grammatically and pragmatically possible antecedents in a preceding clause will be interpreted as coreferential with the candidate that has the same grammatical role" (Smyth 1994:197). Thus, parallelism predicts that in ex. (6), the object pronoun *her* is interpreted as referring to Anne, since both Anne and the pronoun are in object position. In contrast, according to a resolution algorithm based on grammatical role, *her* should be interpreted as referring to the subject of the preceding clause, Lisa (the 'subject strategy').

(6) Lisa saw Anne and Mary saw her too.

According to Kehler (2002), the seemingly conflicting predictions of parallelism and the 'subject strategy' are best viewed as resulting from different kinds of coherence relations between the two sentences. More specifically, he claims that (i) the parallelism strategy is used for pronoun resolution when the first sentence (e.g. *Lisa saw Anne*) and the second sentence (e.g. *Mary saw her too*) are in a parallel relation to each other, and that (ii) the 'subject strategy' is used when the two sentences are related to each other in a more narrative, sequential way. In order for my algorithm - and other pronoun resolution systems that have been proposed in the literature - to be able to take these coherence relations into account, some kind of discourse structure encoding needs to be added to the algorithm itself, such that the algorithm can tell which strategy is likely to be applied in a given context.

However, it is worth noting that algorithms which simply adhere to the 'subject strategy' and do not use parallelism still manage to obtain good empirical coverage (see e.g

[12] Thanks to one of the reviewers for bringing this concern to my attention.

Tetreault (2000)). Thus, it does not seem unreasonable, for the present algorithm, to make use of the 'subject strategy.'

Let us now return to the details of the current version of the algorithm. So far we have considered pre- and postverbal subjects, and preverbal objects. The resolution of a postverbal NP object is less straightforward than the resolution of a preverbal one. In the case of a postverbal object, it is unclear whether the entity already has a peg or whether is new to the discourse. Thus, a search procedure is used to check whether a matching peg is present in the discourse register. The algorithm checks the seven highest pegs[13] to see if one of them is linked to a referent that matches the postverbal NP object. If it finds one, this peg is moved to the second-highest position in the current register. If no peg is found, a new peg is added to the top of the current register, below the subject peg. The details of the matching process are left for future research.

So far, in this section, we have been concerned with SVO and OVS orders. Let us now briefly consider SOV and OSV orders. As mentioned earlier, the immediately preverbal argument in these orders is discourse-old. The initial argument is contrastive, and can be either discourse-old or discourse-new. The algorithm can deal with this 'ambiguity' by a checking mechanism like the one described above. However, after the algorithm has established whether the initial argument is new or old information, it is not clear how the subject and the object should be ranked relative to one another. Preliminary informant judgments suggest that in SOV order, the subject (regardless of whether it is old or new information) is ranked higher than the object, and is thus likely to be interpreted as the referent of a subsequent pronoun. The ranking for OSV order appears to be less stable, and merits further research.[14]

5.2 Pronominal subjects and objects

In this algorithm, the asymmetry in the ranking of subjects over objects also holds for pronominal arguments.[15] When the algorithm encounters a pronominal subject, it looks for the topmost compatible peg in the discourse register and brings it to the top of the current register. However, when the algorithm comes across a pronominal object, it brings the topmost compatible peg from the discourse register to the current register, but does not raise this peg above the peg for the subject. Example (7) illustrates this. According to informant judgments, the gender-neutral subject pronoun *hän* 's/he' (in the third sentence)

[13] The number seven was chosen because human short-term/working memory can contain approximately seven items. See e.g. Miller (1956).

[14] All the examples discussed in this paper involve only bare nouns or pronouns. The issues involved with more complex noun phrases such as *joku nainen* 'some woman' and *toinen kissa* 'another cat' deserve further research, but cannot be addressed here for space reasons. A promising future avenue to pursue for constructions with '(an)other' is offered by some recent work on alternative sets (see e.g. Kruiff-Korbayová and Webber (2001), Bierner and Webber (2000)).

[15] I only discuss pronominal arguments in the canonical SVO order in this section. See Kaiser (to appear) for discussion of preverbal object pronouns.

shows a preference for the preceding subject even when an object pronoun is present (in the second sentence). If this object pronoun raised its peg to very top, we would expect a subsequent pronoun to refer to it.

(7) Pekka keittää kahvia. Liisa katselee häntä samalla kun vesi kiehuu. Hän
Pekka cooks coffee-PART. Liisa looks-at s/he-PART same-time as water-NOM boils. S/he
tykkää Pekasta/???Liisasta.
likes Pekka-ELA/Liisa-ELA.
'Pekka is making coffee. Liisa looks at him while the water is boiling. She likes Pekka/???He likes Liisa.'

Ex. (7) also illustrates the advantage of having a separate discourse register. The distinction between the current register and the discourse register explains why the object pronoun *häntä* 'him/her' in the second sentence cannot refer to the subject of its sentence, Liisa. Pronouns are interpreted by the algorithm as instructions to point to the topmost compatible peg in the discourse register, and in ex. (7), the peg for Liisa is still in the current register and thus it is not a possible antecedent for the pronoun. Another advantage of having both a discourse register and a current register is the ease with which sentences with two pronouns (e.g. *Peter saw John. He kicked him*) can be interpreted (see the end of Section 5.3).

5.3 Anaphoric use of demonstratives

Let us now tackle a sentence where both the subject and object are pronouns, such as *Peter saw John. He kicked him.* In Finnish, such sentences are somewhat marked, due to the existence of another alternative, namely the demonstrative *tämä* 'this,' which can be used as an anaphor (e.g. Hakulinen and Karlsson (1988), Sulkala and Karjalainen (1992)). It has often been noted that this anaphoric demonstrative tends to refer to non-subject arguments (e.g. Saarimaa (1949)). Corpus studies support this observation (Halmari (1994), Kaiser (2000b)). In a two-pronoun sentence, the anaphoric demonstrative is usually used instead of a second pronoun (ex. (8)).

(8) Pekka huomasi Matin pihalla. Tämä tervehti häntä.
Pekka-NOM noticed Matti-ACC yard-ADE. This-NOM greeted s/he-PART.
'Pekka noticed Matti in the yard. He$_{Matti}$ greeted this$_{Pekka}$.'

Given the tendency of *tämä* to refer to postverbal objects, one might hypothesize that, when the algorithm encounters an anaphoric demonstrative, it points to the second highest peg in the discourse register. This generates the correct interpretation for the example above, where *tämä* picks out the object of the preceding sentence. However, ex. (9) below shows that this is insufficient. Here, *tämä* can be used even when only one peg is left in the discourse register.

(9) Pekka huomasi Matin pihalla. Hän tervehti tätä.
Pekka-NOM noticed Matti-ACC yard-ADE. S/he-NOM greeted this-PART.
'Pekka noticed Matti in the yard. He$_{Pekka}$ greeted this$_{Matti}$.'

At the start of the second sentence in (9), the algorithm encounters *hän* 's/he' and pulls the peg for Pekka into the current register. Then the anaphoric demonstrative *tämä* 'this' is encountered – but at this point, only the peg for Matti remains in the discourse register, and it is referred to with *tämä*. One might conclude that the anaphoric demonstrative is *preferably* interpreted as an instruction to point to the second-highest peg but, if no other alternative is available, it can also be interpreted as pointing to the top peg. This approach, however, runs into trouble with ex. (10).

(10) Liisa nukkuu kotona. Hän / ???Tämä on sairas.
Liisa sleeps at-home. S/he / This is sick.
'Liisa is sleeping at home. She is sick.'

Even though the peg for Liisa is the only available peg in the discourse register by the time the algorithm gets to the second sentence, *tämä* 'this' cannot be used to refer to Liisa. Instead of trying to define the anaphoric demonstrative *tämä* by the ranking of the peg it points to, maybe we should treat it as having preference to refer to objects, i.e. to be associated with referents that have a certain grammatical function/semantic role.

Alternatively, we could formulate a more complicated hypothesis as follows: If, at the time the algorithm encounters *tämä*, the current register is empty, then *tämä* refers to the second highest peg in the discourse register as in ex. (8) (as long as the two top pegs in the d.r. both occurred in the preceding sentence). However, when the current register already contains a (particular) peg from the same sentence that also contains the demonstrative *tämä* (ex. (9)), then *tämä* refers to the top peg of the discourse register. In other words, it might be that, in order to correctly resolve *tämä*, the algorithm needs to make use of the two registers simultaneously.[16]

On a more straightforward note, the algorithm works smoothly for 'double-pronoun' sentences of the English type. Consider a sentence such as *Peter saw John. He hit him*, which most people tend to interpret as meaning 'Peter hit John.'[17] When the algorithm encounters the subject pronoun *he* in the second sentence, it pulls the peg for Peter into the current register. When it reaches the object pronoun *him*, the top peg in the discourse register is the peg for John, i.e. the object pronoun is interpreted as referring to John. Thus, due to the presence of the two registers, the algorithm resolves the pronouns successfully, without the need for any additional stipulations concerning the second pronoun.

[16] Thanks to one of the reviewers for bringing this possibility to my attention.

[17] I am assuming neutral intonation.

6 Main vs. Subordinate clauses

An important question that we have sidestepped so far is: What is the size of the utterance? I.e., when are pegs dumped from the current utterance register to the discourse register? This question is especially relevant when it comes to sentences that contain subordinate clauses (see e.g. Kameyama (1998) and Miltsakaki (1999)). Consider the example below (based on a Japanese example given by Miltsakaki (1999). If we treat the subordinate clause as part of the second main clause, then the pronoun *he* in the second sentence is predicted to refer to Peter, the subject of the main clause. However, if we treat the subordinate clause as a separate utterance on a par with a main clause, then we expect the pronoun *he* to refer to the subject of the embedded clause, Alex. The most natural reading of the sentence is one where *he* refers to Peter - in other words, it seems that the subordinate clause does not ’intervene’ between the pronoun and the first main clause. This is what one would expect if it is the case that subordinate clauses are not independent units (see Miltsakaki (1999) for further discussion).

(11) Peter looks a little unhappy. Since Alex is making a great sandcastle, he is jealous.

In fact, on the basis of empirical data from Modern Greek and Japanese, Miltsakaki (1999) argues that, within a Centering Theory approach, one should treat ”the sentence in its traditional sense and not the finite clause” as the appropriate update unit.[18] The distribution of anaphoric demonstratives in Finnish (Kaiser 2000b) provides further evidence for the idea that, for purposes of anaphor resolution, at least some kinds of embedded clauses should be treated as subparts of the main clause. Consider the following example:

(12) Matti sanoi, että Liisa on sairas. Tämä oli saanut flunssan hiihtomatkalla.
Matti$_i$ said that Liisa$_j$ is sick. This$_j$ was gotten flu-ACC ski-trip-ADE
’Matti said that Liisa is sick. She had caught the flu during a skiing trip.

As this example shows, in addition to referring to objects (Section 5), the anaphoric demonstrative *tämä* can also be used to refer to an embedded subject - especially when a ‘competitor’ antecedent is present in the main clause. If a subordinate clause behaves just like a main clause, the regular pronoun *hän* should be used to refer to the subjects of all kinds of sentences equally, regardless of whether they are matrix or subordinate. The preference to use *tämä* for referents in subordinate clauses thus suggests that, for purposes of anaphora resolution, subordinate clauses are best treated as subparts of the main clause.

We can encode this in the algorithm by introducing a partition in the current register, such that subordinate sentences have their own separate current register alongside the

[18] It is worth noting that in addition to raising the question about utterance size, subordinate clauses also highlight the need for real dynamics. Consider a sentence such as *If Pekka noticed someone in the yard, he greeted him.* In terms of dynamic semantics, *someone* is not accessible because it is in a conditional context. In terms of my algorithm, we could say that *someone* evokes a peg which is temporarily in the discourse register. Thanks to one of the reviewers for bringing this to my attention.

current register of the main sentence. E.g., for ex. (12), when the algorithm has processed the main clause, it dumps the current register of the main clause into the discourse register, and then after processing the subordinate clause, it dumps its current register into the discourse register below that of the main clause, as if this second current register were an object in some sense.

7 Conclusion

In this paper, I have presented a preliminary referent-tracking and pronoun resolution system for Finnish, an articleless, free-word-order language that poses a challenge for entity-tracking systems designed for languages like English or German. The algorithm described in this paper uses the word order patterns of Finnish to distinguish 'old' and 'new' referents. In addition, by mapping the entities in a discourse onto pegs that are ranked with respect to their salience, the algorithm aims to function as a pronoun resolution system.

Many issues remain open for future research, including the interpretation of plural pronouns, the role of intonation and the treatment of more complex noun phrases. Moreover, so far, the algorithm has only been tested on very small amounts of text. In the future, it should be tested on larger corpora and also compared to/combined with existing algorithms, in order to determine how to best tackle referent tracking and resolution in a language of this type.

References

Bierner, G. and B. Webber (2000). Inference through Alternative-set Semantics. *Journal of Language and Computation 1*, 259–274.

Chesterman, A. (1991). *On definiteness.* Cambridge University Press.

Groenendijk, J., M. Stokhof, and F. Veltman (1996). Coreference and Modality. In S. Lappin (Ed.), *The Handbook of Contemporary Semantic Theory*, pp. 179–213. Blackwell.

Grosz, B., A. Joshi, and S. Weinstein (1995). Centering: A Framework for Modelling the Local Coherence of Discourse. Technical report, Instute for Research in Cognitive Science, University of Pennsylvania.

Hakulinen, A. and F. Karlsson (1979). Kvantitatiivinen tutkimus suomen morfosyntaksista ja sen tekstuaalista tekijöistä. Ms., University of Turku.

Hakulinen, A. and F. Karlsson (1988). *Nykysuomen lauseoppia.* Suomalaisen Kirjallisuuden Seura.

Halmari, H. (1994). On accessibility and coreference. *Nordic Journal of Linguistics 17*, 35–59.

Hintikka, J. and G. Sandu (1997). Game-Theoretical Semantics. In J. van Benthem and A. ter Meulen (Eds.), *Handbook of Logic and Language*, pp. 361–410. Elsevier Science.

Hoffman, B. (1998). Word Order, Information Structure and Centering in Turkish. In M. Walker, A. Joshi, and E. Prince (Eds.), *Centering Theory in Discourse*, pp. 251–272. Clarendon Press.

Hudson-D'Zmura, S. and M. Tanenhaus (1998). Assiging Antecedents to Ambiguous Pronouns: The Role of the Center of Attention as the Default Assignment. In A. J. M.A. Walker and E. Prince (Eds.), *Centering Theory in Discourse*, pp. 199–226. Clarendon Press.

Kaiser, E. (2000a). The discourse functions and syntax of OSV word order in Finnish. In *Proceedings of the 36th Annual Meeting of the Chicago Linguistics Society.*

Kaiser, E. (2000b). Pronouns and demonstratives in Finnish: Indicators of Referent Salience. In *Proceedings of the Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC).*

Kaiser, E. (to appear). Word order, grammatical function, and referential form: On the patterns of anaphoric reference in Finnish. In *Proceedings of the 19th Scandinavian Conference of Linguistics.*

Kameyama, M. (1998). Intrasentential Centering: A case Study. In A. J. M.A. Walker and E. Prince (Eds.), *Centering Theory in Discourse*, pp. 98–112. Clarendon Press.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar.* CSLI.

Kruiff-Korbayová, I. and B. Webber (2001). Concession, Implicature and Alternative Sets. In *Proceedings of the International Workshop on Computational Semantics (IWCS-4).*

Kruijff-Korbayová, I. (1997). Czech Noun Phrases in File Change Semantics. In G. K. A. Drewery and R. Zuber (Eds.), *Proceedings of the Student Session at the 10th ESSLLI*, pp. 107–118.

Laury, R. (1997). *Demonstratives in Interaction - The Emergence of a Definite Article in Finnish.* John Benjamins.

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review 3*, 81–97.

Miltsakaki, E. (1999). Locating Topics in Text Processing. In *Proceedings of Computational Linguistics in the Netherlands (CLIN'99).*

Prasad, R. and M. Strube (2000). Discourse Salience and Pronoun Resolution in Hindi. *UPenn Working Papers in Linguistics 6*, 189–208.

Prince, E. (1992). The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann (Eds.), *Discourse description: diverse analyses of a fund raising text*, pp. 295–325. John Benjamins.

Rambow, O. (1993). Pragmatic Aspects of Scrambling and Topicalization in German. Paper presented at the Workshop on Naturally-Occurring Discourse, IRCS, University of Pennsylvania.

Saarimaa, E. (1949). Kielemme käytäntö. Pronominivirheistä. *Virittäjä 49*, 250–257.

Smyth, R. (1994). Grammatical Determinants of Ambiguous Pronoun Resolution. *Journal of Psycholinguistic Research 23*, 197–229.

Strube, M. (1998). Never Look Back: An Alternative to Centering. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the ACL*.

Strube, M. and U. Hahn (1996). Functional Centering. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*.

Sulkala, H. and M. Karjalainen (1992). *Finnish*. Routledge.

Tetreault, J. (submitted). A Corpus-Based Evaluation of Centering and Pronoun Resolution. *Computational Linguistics*.

Turan, U. D. (1998). Ranking Forward-Looking Centers in Turkish: Universal and Language-Specific Properties. In M. Walker, A. Joshi, and E. Prince (Eds.), *Centering Theory in Discourse*, pp. 139–160. Clarendon Press.

Vallduví, E. and M. Vilkuna (1998). On rheme and kontrast. In P. Culicover and L. McNally (Eds.), *The Limits of Syntax Syntax and Semantics 29*, pp. 79–108. Academic Press.

Vilkuna, M. (1995). Discourse Configurationality in Finnish. In K. E. Kiss (Ed.), *Discourse Configurational Languages*, pp. 244–268. Oxford University Press.

Walker, M. (1998). Centering: Anaphora Resolution and Discourse Structure. In A. J. M.A. Walker and E. Prince (Eds.), *Centering Theory in Discourse*, pp. 401–435. Clarendon Press.

Walker, M., A. Joshi, and E. Prince (1998). *Centering Theory in Discourse*. Clarendon Press.