Translating into Free Word Order Languages

Beryl Hoffman
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, U.K.
hoffman@cogsci.ed.ac.uk

Abstract

In this paper, I discuss machine translation of English text into a relatively "free" word order language, specifically Turkish. I present algorithms that use contextual information to determine what the topic and the focus of each sentence should be, in order to generate the contextually appropriate word orders in the target language.

1 Introduction

Languages such as Catalan, Czech, Finnish, German, Hindi, Hungarian, Japanese, Polish, Russian, Turkish, etc. have much freer word order than English. For example, all six permutations of a transitive sentence are grammatical in Turkish (although SOV is the most common). When we translate an English text into a "free" word order language, we are faced with a choice between many different word orders that are all syntactically grammatical but are not all felicitous or contextually appropriate. In this paper, I discuss machine translation (MT) of English text into Turkish and concentrate on how to generate the appropriate word order in the target language based on contextual information.

The most comprehensive project of this type is presented in (Stys/Zemke, 1995) for MT into Polish. They use the referential form and repeated mention of items in the English text in order to predict the salience of discourse entities and order the Polish sentence according to this salience ranking. They also rely on statistical data, choosing the most frequently used word orders. I argue for a more generative approach: a particular information structure (IS) can be determined from the contextual information and then can be used to generate the felicitous word order. This paper concentrates on how to determine the IS from contextual information using centering, old vs. new information, and contrastiveness. (Hajičová/etal, 1993; Steinberger, 1994) present approaches that determine the IS by using cues such as word order, definiteness, and complement semantic types (e.g. temporal adjuncts vs arguments) in the source language, English. I believe that we cannot rely upon cues in the source language in order to determine the IS of the translated text. Instead, I use contextual information in the target language to determine the IS of sentences in the target language.

In section 2, I discuss the Information Structure, and specifically the topic and the focus in naturally occurring Turkish data. Then, in section 3, I present algorithms for determining the topic and the focus, and show that we can generate contextually appropriate word orders in Turkish using these algorithms in a simple MT implementation.

2 Information Structure

In the Information Structure (IS) that I use for Turkish, a sentence is first divided into a topic and a comment. The topic is the main element that the sentence is about, and the comment is the information conveyed about this topic. Within the comment, we find the focus, the most information-bearing constituent in the sentence, and the ground, the rest of the sentence. The focus is the new or important information in the sentence and receives prosodic prominence in speech.

In Turkish, the pragmatic function of topic is assigned to the sentence-initial position and the focus to the immediately preverbal position, following (Erguvanlı, 1984). The rest of the sentence forms the ground.

In (Hoffman, 1995; Hoffman, 1995b), I show that the information structure components of topic and focus can be successfully used in generating the context-appropriate answer to database queries. Determining the topic and focus is fairly easy in the context of a simple question, however it is much more complicated in a text. In the fol-

The Cb in SOV sentences.		
Cb = Subject	14	(47%)
Cb = Object	6	(20%)
Cb = Subj or Obj?	6	(20%)
Cb = Subj or Other Obj?	0	(0%)
No Cb	4	(13%)
TOTAL	30	

The Cb in OSV sentences.		
Cb = Subject	4	(13%)
Cb = Object	16	(53%)
Cb = Subj or Obj?	6	(20%)
Cb = Subj or Other Obj?	2	(7%)
No Cb	2	(7%)
TOTAL	30	

Figure 1: The Cb in SOV and OSV Sentences.

lowing sections, I will describe the characteristics of topic, focus, and ground components of the IS in naturally occurring texts analyzed in (Hoffman, 1995b) and allude to possible algorithms for determining them. The algorithms will then be spelled out in section 3.

An example text from the corpus[1] is shown below. The noncanonical OSV word order in (1)b is contextually appropriate because the object pronoun is a discourse-old topic that links the sentence to the previous context, and the subject, "your father", is a discourse-new focus that is being contrasted with other relatives. *Discourse-old* entities are those that were previously mentioned in the discourse while *discourse-new* entities are those that were not (Prince, 1992).

(1) a.
Bu defteri de çok sevdim ben.
This notebk-acc too much like-pst-1S I.
'As for this notebook, I like it very much.'

b.
Bunu da baban mı verdi? (OSV)
This-Acc too father-2S Quest give-Past?
'Did your FATHER give this to you?'
(CHILDES 1ba.cha)

Many people have suggested that "free" word order languages order information from old to new information. However, the Old-to-New ordering principle is a generalization to which exceptions can be found. I believe that the order in which speakers place old vs. new items in a sentence reflects the information structures that are available to the speakers. The ordering is actually the Topic followed by the Focus. The Topic tends to be discourse-old information and the focus discourse-new. However, it is possible to have a discourse-NEW topic and a discourse-OLD focus, as we will see in the following sections, which explains the exceptions to the Old-To-New ordering principle.

[1]The data was collected from transcribed conversations, contemporary novels, and adult speech from the CHILDES corpus.

2.1 Topic

Although humans can intuitively determine what the topic of a sentence is, the traditional definition (what the sentence is about) is too vague to be implemented in a computational system. I propose heuristics based on familiarity and salience to determine discourse-old sentence topics, and heuristics based on grammatical relations for discourse-new topics. Speakers can shift to a new topic at the start of a new discourse segment, as in (2)a. Or they can continue talking about the same discourse-old topic, as in (2)b.

(2) a. [Mary]$_T$ went to the bookstore.
b. [She]$_T$ bought a new book on linguistics.

A discourse-old topic often serves to link the sentence to the previous context by evoking a familiar and salient discourse entity. Centering Theory (Grosz/etal, 1995) provides a measure of saliency based on the observations that salient discourse entities are often mentioned repeatedly within a discourse segment and are often realized as pronouns. (Turan, 1995) provides a comprehensive study of null and overt subjects in Turkish using Centering Theory, and I investigate the interaction between word order and Centering in Turkish in (Hoffman, 1996).

In the Centering Algorithm, each utterance in a discourse is associated with a ranked list of discourse entities called the forward-looking centers (Cf list) that contains every discourse entity that is realized in that utterance. The Cf list is usually ranked according to a hierarchy of grammatical relations, e.g. subjects are assumed to be more salient than objects. The backward looking center (Cb) is the most salient member of the Cf list that links the current utterance to the previous utterance. The Cb of an utterance is defined as the highest ranked element of the previous utterance's Cf list that also occurs in the current utterance. If there is a pronoun in the sentence, it is likely to be the Cb. As we will see, the Cb has much in common with a sentence-topic.

	S-init SOV,OSV	IPV SOV,OSV	Post-V OVS, SVO
Discourse-Old	55 (85%)	43 (67%)	56 (93%)
Inferrable	8 (13%)	10 (16%)	4 (7%)
D-New, Hearer-Old	1 (2%)	1 (2%)	0
⋆ D-New, Hearer-New	0	10 (15%)	0
TOTAL	64	64	60

Figure 2: Given/New Status in Different Sentence Positions

The Cb analyses of the canonical SOV and the noncanonical OSV word orders in Turkish are summarized in Figure 1 (forthcoming study in (Hoffman, 1996)). As expected, the subject is often the Cb in the SOV sentences. However, in the OSV sentences, the object, not the subject, is most often the Cb of the utterance. A comparison of the 20 discourses in the first two rows[2] of the tables in Figure 1 using the chi-square test shows that the association between sentence-position and Cb is statistically significant ($\chi^2 = 10.10, \rho < 0.001$).[3] Thus, the Cb, when it is not dropped, is often placed in the sentence initial topic position in Turkish regardless of whether it is the subject or the object of the sentence. The intuitive reason for this is that speakers want to form a coherent discourse by immediately linking each sentence to the previous ones by placing the Cb and discourse-old topic in the sentence-initial position.

There are also situations where no Cb or discourse-old topic can be found. Then, a discourse-new topic can be placed in the sentence-initial position to start a new discourse segment. Discourse-new topics are often subjects or situation-setting adverbs (e.g. yesterday, in the morning, in the garden) in Turkish.

2.2 Focus

The term focus has been used with many different meanings. Focusing is often associated with new information, but it is well-known that old information, for example pronouns, can be focused as well. I think part of the confusion lies in the distinction between contrastive and presentational focus. Focusing discourse-new information is often called presentational or informational focus as shown in (3)a. Broad/wide focus (focus projection) is also possible where the rightmost element in the phrase is accented, but the whole phrase is in focus. However, we can also use focusing in order to contrast one item with another, and in this case the focus can be discourse-old or discourse-new, e.g. (3)b.

(3) a. What did Mary do this summer?
She [wandered around TURKEY]$_F$.
b. It wasn't [ME]$_F$ - It was [HER]$_F$.

(Vallduví, 1992) defines focus as the most information-bearing constituent, and this definition encompasses both contrastive and presentational focusing. I use this definition of focus as well. However, as will see, we still need two different algorithms in order to determine which items are in focus in the target sentence in MT. We must check to see if they are discourse-new information as well as checking if they are being contrasted with another item in the discourse model.

In Turkish, items that are presentationally or contrastively focused are placed in the immediately preverbal (IPV) position and receive the primary accent of the phrase.[4] As seen in Figure 2, brand-new discourse entities are found in the IPV position, but never in other positions in the sentence in my Turkish corpus. The distribution of brand-new (the starred line of the table) versus discourse-old information (the rest of the table[5]) is statistically significant, ($\chi^2 = 10.847, \rho < .001$). This supports the association of discourse-new focus with the IPV position.

[2]The centering analysis is inconclusive in some cases because the subject and the object in the sentence are realized with the same referential form (e.g. both as overt pronouns or as full NPs).

[3]Alternatively, using the canonical SOV sentences as the expected frequencies, the observed frequencies for the noncanonical OSV sentences significantly diverge from the expected frequencies ($\chi^2 = 8.8, \rho < 0.005$).

[4]Some languages such as Greek and Russian treat presentational and contrastive focus differently in word order.

[5]*Inferrables* refer to entities that the hearer can easily accommodate based on entities already in the discourse model or the situation. *Hearer-old* entities are well-known to the speaker and hearer but not necessarily mentioned in the prior discourse (Prince, 1992). They both behave like discourse-old entities.

However, as can be seen in Figure 2, most of the focused subjects in the OSV sentences in my corpus were actually discourse-old information. Discourse-old entities that occur in the IPV position are contrastively focused. In (Rooth, 1985)'s alternative-set theory, a contrastively focused item is interpreted by constructing a set of alternatives from which the focused item must be distinguished. Generalizing from his work, we can determine whether an entity *should* be contrastively focused by seeing if we can construct an alternative set from the discourse model.

2.3 Ground

Those items that do not play a role in IS of the sentence as the topic or the focus form the ground of the sentence. In Turkish, discourse-old information that is not the topic or focus can be

(4) a. dropped,
    b. postposed to the right of the verb,
    c. or placed unstressed between the topic and the focus.

Postposing plays a backgrounding function in Turkish, and it is very common. Often, speakers will drop only those items that are very salient (e.g. mentioned just in the previous sentence) and postpose the rest of the discourse-old items. However, the conditions for dropping arguments can be very complex. (Turan, 1995) shows that there are semantic considerations; for instance, generic objects are often dropped, but specific objects are often realized as overt pronouns and fronted. Thus, the conditions governing dropping and postposing are areas that require more research.

3 The Implementation

In order to simplify the MT implementation, I concentrate on translating short and simple English texts into Turkish, using an interlingua representation where concepts in the semantic representation map onto at most one word in the English or Turkish lexicons. The translation proceeds sentence by sentence (leaving aside questions of aggregation, etc.), but contextual information is used during the incremental generation of the target text. These simplifications allow me to test out the algorithms for determining the topic and the focus presented in this section.

In the implementation, first, an English sentence is parsed with a Combinatory Categorial Grammar, CCG, (Steedman, 1985). The semantic representation is then sent to the sentence planner for Turkish. The Sentence Planner uses the algorithms in the following subsections to determine the topic, focus, and ground from the given semantic representation and the discourse model. Then, the sentence planner sends the semantic representation and the information structure it has determined to the sentence realization component for Turkish. This component consists of a head-driven bottom up generation algorithm that uses the semantic as well as the information structure features given by the planner to choose an appropriate head in the lexicon. The grammar used for the generation of Turkish is a lexicalist formalism called Multiset-CCG (Hoffman, 1995; Hoffman, 1995b), an extension of CCGs. Multiset-CCG was developed in order to capture formal and descriptive properties of "free" and restricted word order in simple and complex sentences (with discontinuous constituents and long distance dependencies). Multiset-CCG captures the context-dependent meaning of word order in Turkish by compositionally deriving the predicate-argument structure and the information structure of a sentence in parallel.

The following sections describe the algorithms used by the sentence planner to determine the IS of the Turkish sentence, given the semantic representation of a parsed English sentence.

3.1 The Topic Algorithm

As each sentence is translated, we update the discourse model, and keep track of the forward looking centers list (Cflist) of the last processed sentence. This is simply a list of all the discourse enities realized in that sentence ranked according to the theta-role hierarchy found in the semantic representation. Thus, the Cf list for the representation *give(Pat,Chris,book)* is the ranked list **[Pat,Chris,book]**, where the subject is assumed to be more salient than the objects.

Given the semantic representation for the sentence, the discourse model of the text processed so far, and the ranked Cf lists of the current and previous sentences in the discourse, the following algorithm determines the topic of the sentence. First, the algorithm tries to choose the most salient discourse-old entity as the sentence topic.[6] If there is no discourse-old entity realized in the sentence, then a situation-setting adverb or the subject is chosen as the discourse-new topic.

1. Compare the current Cf list with the previous sentence's Cf list and choose the first item that is a member of both of the ranked lists (the Cb).

[6] (Stys/Zemke, 1995) use the saliency ranking to order the whole sentence in Polish. However, I believe that there is a distinct notion of topic and focus in Turkish.

2. If 1 fails: Choose the first item in the current sentence's Cf list that is discourse-old (i.e. is already in the discourse model).

3. If 2 fails: If there is a situation-setting adverb in the semantic representation (i.e. a predicate modifying the main event in representation), choose it as the discourse-new topic.

4. If 3 fails: choose the first item in the Cf list (i.e. the subject) as the discourse-new topic.

Note that the determination of the sentence topic is distinct from the question of how to realize the salient Cb/topic (e.g. as a dropped or overt pronoun or full NP). In the MT domain, this can be determined by the referential form in the source text. This trick can also be used for accommodating inferrable or hearer-old entities that behave as if they are discourse-old even though they are literally discourse-new. If an item that is not in the discourse model is nonetheless realized as a definite NP in the source text, the speaker is treating the entity as discourse-old. This is very similar to (Stys/Zemke, 1995)'s MT system which uses the referential form in the source text to predict the topicality of a phrase in the target text.

3.2 The Focus Algorithm

Given the rest of the semantic representation for the sentence and the discourse model of the text processed so far, the following algorithm determines the focus of the sentence. The first step is to determine presentational focusing of discourse-new information. Note that the focus, unlike the topic, can contain more than one element; this allows broad focus as well as narrow focusing. If there is no discourse-new information, the second step in the algorithm allows contrastive focusing of discourse-old information. In order to construct the alternative sets, a small knowledge base is used to determine the semantic type (agent, object, or event) of the entities in the discourse model.

1. If there are any discourse-new entities (i.e. not in the discourse model) in the sentence, put their semantic representations into focus.

2. Else for each discourse entity realized in the sentence,
   (a) Look up its semantic type in the KB and construct an alternative set that consists of all objects of that type in the discourse model,
   (b) If the constructed alternative set is not empty, put the discourse entity's semantic representation into the focus.

Once the topic and focus are determined, the remainder of the semantic representation is assigned as the ground. For now, items in the ground are either generated in between the topic and the focus or post-posed behind the verb as backgrounded information. Further research is needed to disambiguate the use of the two possible word orders.

Further research is also needed on the exact role of verbs in the IS. Verbs can be in the focus or the ground in Turkish; this cannot be seen in the word order, but it is distinguished by sentential stress for narrow focus readings. The algorithm above works for verbs since I place events that are realized as verbs in the sentence into the discourse model as well. However, verbs are usually not in focus unless they are surprising or contrastive or in a discourse-initial context. Thus, the algorithm needs to be extended to accommodate discourse-new verbs that are nonetheless expected in some way into the ground component. In addition, verbs often participate in broad focus readings, and further research is needed to account for the observation that broad focus readings are only available in canonical word orders.

3.3 Examples

The English text in (5) is translated using the word orders in (6) following the algorithms given above. In (6), the numbers following T and F indicate the step in the respective algorithm which determined the topic or focus for that sentence. Note that the inappropriate word orders (indicated by #) cannot be generated by the algorithm.

(5) a. Pat will meet Chris today.
    b. There is a talk at four.
    c. Chris is giving the talk.
    d. Pat cannot come.

(6) a.
    Bugün Pat Chris'le buluşacak. (AdvSOV)
    Today Pat Chris-with meet-fut. (T:3,F:1)

    b.
    Dörtde bir konuşma var. (AdvSV,#SAdvV)
    Four-Loc one talk exist. (T:3,F:1)

    c. Konuşmayı Chris veriyor. (OSV,#SOV)
    Talk-Acc Chris give-Prog. (T:1,F:2)

    d.
    Pat gelemiyecek. (SV,#VS)
    Pat come-Neg-Fut. (T:2,F:1 for the verb)

The algorithms can also utilize long distance scrambling in Turkish, i.e. constructions where an element of an embedded clause has been ex-

tracted and scrambled into the matrix clause in order to play a role in the IS of the matrix clause. For example the b sentence in the following text is translated using long distance scrambling because "the talk" is the Cb of the utterance and therefore the best sentence topic, even though it is the argument of an embedded clause.

(7) a. There is a talk at four.
b. Pat thinks that Chris will give the talk.

(8) a. Dörtde bir konuşma var. (AdvSV)
Four-Loc one talk exist.

b.
Konuşmayı$_i$ Pat [Chris'in e_i vereceğini]
Talk-Acc$_i$ Pat [Chris-gen e_i give-ger-3s-acc]

sanıyor. ($O_2S_1[S_2V_2]V_1$)
think-Prog. (T:1,F:1)

4 Conclusions

In the machine translation task from English into a "free" word order language, it is crucial to choose the contextually appropriate word order in the target language. In this paper, I discussed how to determine the appropriate word order using contextual information in translating into Turkish. I presented algorithms for determining the topic and the focus of the sentence. These algorithms are sensitive to whether the information is old or new in the discourse model (incrementally constructed from the translated text); whether they refer to salient entities (using Centering Theory); and whether they can be contrasted with other entities in the discourse model. Once the information structure for a semantic representation is constructed using these algorithms, the sentence with the contextually appropriate word order is generated in the target language using Multiset CCG, a grammar which integrates syntax and information structure.

References

Eser Emine Erguvanlı. 1984. *The Function of Word Order in Turkish Grammar.* University of California Press.

Barbara Grosz and Aravind K. Joshi and Scott Weinstein. 1995. Centering: A Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics.*

Hajičová, Eva, Petr Sgall, and Hana Skoumalová. 1993. Identifying Topic and Focus by an Automatic Procedure. *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics.*

Beryl Hoffman. 1995. Integrating Free Word Order Syntax and Information Structure. *Proceedings of the European Association for Computational Linguistics (EACL).*

Beryl Hoffman. 1995. *The Computational Analysis of the Syntax and Interpretation of "Free" Word Order in Turkish.* Ph.D. dissertation. IRCS Tech Report 95-17. Dept. of Computer and Information Science. University of Pennsylvania.

Beryl Hoffman. to appear 1996. Word Order, Information Structure, and Centering in Turkish. *Centering in Discourse.* eds. Ellen Prince, Aravind Joshi, and Marilyn Walker. Oxford University Press.

Ellen F. Prince. The ZPG Letter: Subjects, Definiteness and Information Status. *Discourse description: diverse analyses of a fund raising text.* eds. Thompson, S. and Mann, W. Philadelphia: John Benjamins B.V. pp.295-325. 1992.

Mats Rooth. 1985. Association with Focus. Ph.D. Dissertation. University of Massachusetts. Amherst.

Mark Steedman. 1985. Dependencies and coordination in the grammar of Dutch and English, *Language*, 61:523-568.

Ralf Steinberger. 1994. Treating Free Word Order in Machine Translation. *Coling*, Kyoto, Japan.

Malgorzata E. Stys and Stefan S. Zemke. 1995. Incorporating Discourse Aspects in English-Polish MT: Towards Robust Implementation. *Recent Advances in NLP.*

Ümit Turan. 1995. *Null vs. Overt Subjects in Turkish Discourse: A Centering Analysis.* University of Pennsylvania, Linguistics Ph.D. dissertation.

Enric Vallduví. 1992. *The Informational Component.* New York: Garland.