SABINE KIRCHMEIER-ANDERSEN

Coordination in Eurotra

Abstract

The treatment of coordination in a multilingual MT-system as EUROTRA poses two major problems:

1. to provide an algorithm for the monolingual analysis of coordinate structures in accordance with the general linguistic theory for EUROTRA which can treat basic coordination and the coordination of incomplete constituents in the surface analysis and perform the mapping of these constructions onto the deeper levels.
2. to establish a semantic feature system, that captures the meaning of the conjunctions in order to facilitate their translation.

Since research in this area within Eurotra on a multilingual basis is still ongoing, what will be presented is the present state of the art, which is a fairly complete theory for basic coordination as well as some ideas for the analysis of the more complex cases of coordination involving gapping and movement.

1 Introduction

It is well known that coordination is one of the most prevalent and complex constructions in European languages which causes great difficulties in all syntactic formalisms. At present in Eurotra, not all aspects of coordination are covered. The current implementations of the grammars of the monolingual modules handle basic coordination of alike constituents, but not special cases like coordination of unlike or incomplete constructions. Coordination is still a topic of ongoing research both from a monolingual and a contrastive point of view.

In the first part of this paper, I would like to illustrate how coordination is handled in Eurotra in analysis, in transfer and in generation and I will comment on some of the universal mechanisms or constraints, which we exploit in order to avoid overgeneration and to facilitate the translation of coordinated structures. (For information on the Eurotra linguistic theory as well as the overall objectives of the project see Perschke 1986, Jaspaert 1986, Arnold 1986 and Raw et al. 1989). In the second part I will go into two of the difficult areas, that we are working on at the moment.

Proceedings of NODALIDA 1989, pages 191-203

2 Basic Coordination—an Example

Basic coordination is defined as coordination of constituents with the same category, i.e. "John and Mary", "happy or sad" etc.

The following section gives a short description of the translation of a sentence containing a coordinated structure from Danish into English. The following example is used:

(1) Både USA, Japan og Kina viser interesse for EF

both usa, japan and china show interest for eec
(literal translation)

2.1 Analysis

The figures 1 and 2 show the structural representation of the sentence (1) at two of the syntactic levels in the Eurotra theory: ECS—the constituent structure,

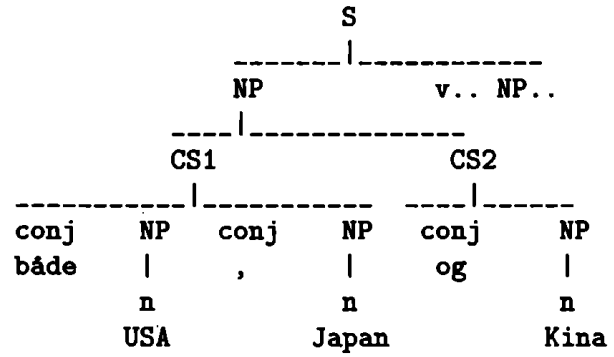


Figure 1: ECS

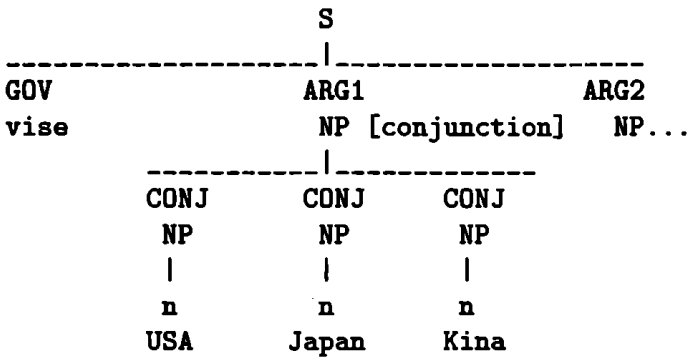


Figure 2: IS

and IS—the deep syntactic structure. Between these levels we have a third level, ERS—the surface syntactic structure, but with respect to coordination the ERS and IS representations are quite similar so ERS is left out in these examples.

Comparing the two structural representations in figure 1 and figure 2, we can note that the intermediate nodes at ECS, "cs1" and "cs2", have been deleted at IS. Their main function is to group the conjuncts and the conjunctions together in the constituent analysis at ECS and to prevent other categories from entering the coordinate structure during the parsing process. The representation at ECS is in some respects similar to the one presented by Gazdar et al. (1985). GPSG uses a constituent similar to the cs2 node for all parts of the coordinated structure. We have introduced a more complex constituent, cs1, for the first two conjuncts in order to speed up the parsing process.

The IS level, where the constituents are defined as dependency structures consisting of a governor and a number of arguments and modifiers, is characterized by the canonical ordering of these constituents. Since coordination is not a dependency relation, the surface order of the constituents in coordinate structures is not changed.

The conjunction, the comma and the preconjunction "både" are deleted at IS. Their semantic content is preserved as a feature "[conjunction]" on the top node of the coordinated structure.

2.2 Transfer

In the transfer phase, the Danish-English transfer module translates the lexical values of the conjuncts into English. This is practically all that happens. The rest of the structure is translated by a default mechanism, that simply tranfers it into the target language IS structure without any changes (figure 3).

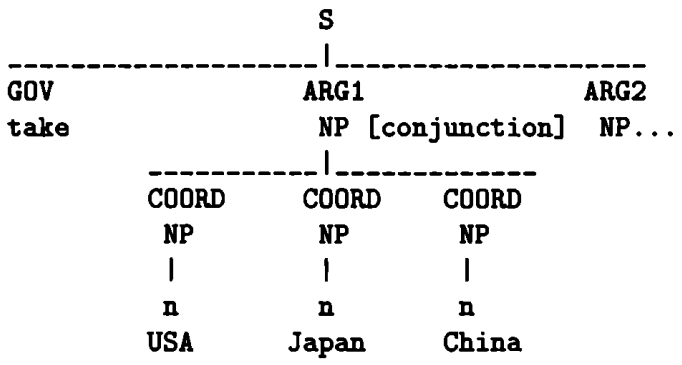


Figure 3: IS

2.3 English Synthesis

The last step of the translation process which will concern us here, is performed during generation in the ECS grammar of the target language, where the English conjunctions are inserted on the basis of the semantic features computed in the analysis. This gives us a structural representation like the one in figure 4.

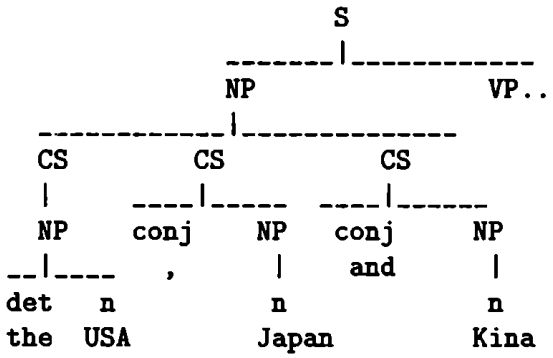


Figure 4: ECS

The final result is the following sentence:

(2) The USA, Japan and China take interest in the EEC

At ECS in generation the "cs" nodes have reappeared. They make it easier to insert the conjunctions in the right places. In synthesis we do not need to distinguish between "cs1" and "cs2", because we know from the analysis, exactly which constituents are to be conjoined.

An equivalent to the preconjunction "både" has not been created and this is part of the justification for the featurization of the conjunctions, since the occurrence of these preconjunctions show monolingual variations. In English "both" is considered a strictly binary conjunction, probably because of the homonymy with the quantifier wich corresponds to "begge" in Danish. In Danish "både" may occur with any number of conjuncts.

We can also note, that the comma has been translated as a comma and not as a lexical conjunction as in (3):

(3) Japan and the USA and China take interest in the EEC

This is a matter of style and taste. Like a joker in a game of cards, the comma can take the place of any conjunction in a coordinate structure, provided that there are at least three conjuncts and that the comma does not appear with the first or the last conjuncts (except in enumerations). If we allowed such variations we would get a multiple output, and since the output of a machine translation system perferably should be one solution, we have chosen the one in (2).

3 Universal Constraints

In order to perform the translation described above, we have relied on a number of general properties, that seem to be the same for all coordinated structures.

3.1 Bar Level Constraints

In order to be able to coordinate all categories, we want to underspecify the category value. The advantage is of course that we can use only one basic set of rules to handle all cases of coordination of alike constituents:

```
1.      X -> cs1[X]  *cs2[X]

2.      cs1[X] -> (preconj) X conj X
3.      cs2[X] ->  conj X
```

X is a variable which is instantiated with the category value of the conjunct and percolated to the top node of the coordinated structure. Using unification we can ensure, that X in every rule contains the same category. These very general rules, however, tend to coordinate categories at every bar level thus performing an anlysis which is obviously wrong. Consider cases as (4):

```
    *  the man   and    boy
(4)      NP             N
        bar=1          bar=0
```

Consequently, we allow only coordination of constituents at the same bar level and not at bar level zero as in (5). A similar approach has been developed by Nirenburg (1989).

```
(5)      *   NP  bar=1
         ____|__________
         det           n   bar=0
         the           |
                      CS1
                 _____|________
                 n    conj    n  bar=0
                 USA  and     China
```

3.2 Binary and Iterative Coordination

For the insertion of the correct conjunctions in synthesis it is important to calculate whether the coordinate construction consists of two or more than two elements, as demonstrated for "both" in (1) and (2). A coordinate structure is binary if it consists of only two conjuncts as in (6):

(6) Both X and Y

Iterative coordinate structures consist of more than two conjuncts as in (7):

(7) X and Y and Z and .....

The occurrence of preconjunctions in binary and iterative coordination is language specific as shown in the translation example above.

3.3 Hierarchy

If we want to introduce the concept of hierarchy for coordinated structures, it entails that our rules have to be recursive. This means that we must be able to create deep representations as in figure 5 as well as the flat ones presented in figure 1 and figure 4.

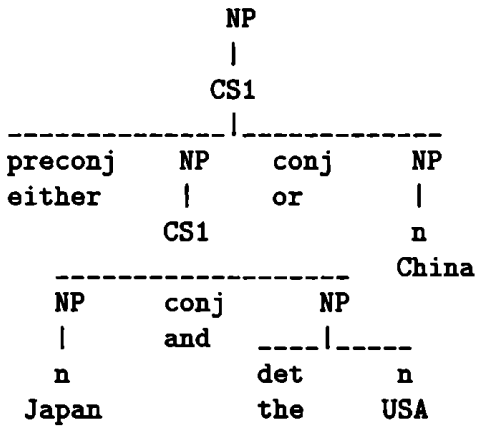


Figure 5: ECS

This structure is hierarchical in the sense that one of the conjuncts is a coordinated structure itself. Allowing this type of recursion without restrictions would cause heavy overgeneration at ECS. The number of possible structures grows fast with the number of conjuncts. 2 conjuncts = 1, 3 conjuncts = 3, 4 conjuncts = 11 etc. Again we prefer only one solution. The constraint that we pose on these structures is simply that coordinate structures may not immediately dominate coordinate structures of the same type (see 3.4 below). In the Eurotra formalism this is expressed in terms of filter or killer rules which delete the undesired structure.

3.4 Cooccurrence Restrictions

Finally, we have to pose cooccurrence restrictions on the conjunctions both with regard to their semantic type and their paradigmatic distribution. The way in which cooccurrence restrictions are expressed is a monolingual matter. For European languages, we can distinguish 5 basic universal types of coordination.

Coordination	Type
både – og, og, samt	conjunction
enten – eller	disjunction
hverken – eller	negation
men	adversion
,	enumeration

The semantic restrictions imply that we cannot have a coordination like (8) on the same hierarchical level.

(8) * både X eller Y

The positional restrictions for the conjunctions operate with 3 positions: initial, non-initial and final. In binary coordination only the initial and final positions are used. In iterative coordination a theoretically infinite number of conjunctions may additionally appear in non-intial position.

Initial	non-initial	final	Example
enten	,/eller	eller	enten A,B eller C
Ø	,/eller	eller	A,B eller C
både	,/og	og/samt	både A,B og C
Ø	,/og	og/samt	A,B og C
hverken	,/eller	eller	hverken A,B eller C

Since coordinate structures consisting of alike constituents are the most frequent ones, the rules and constraints discussed above can handle many of the cases of coordination occurring in the text types we are working with. However, there are still a number of cases where this basic treatment is not sufficient, some of which will be treated in the second part of this paper.

4 Complex Cases of Coordination

4.1 Coordination of Unlike Constituents

The question is now how to integrate the coordination of unlike constituents into the system sketched above.

By coordination of unlike constituents we mean coordinated structures consisting of different categories as in (1)–(5):

(1) Hun sang smukt og med høj stemme
adv prep + sub

(2) Han var glad og i godt humør
adj prep + sub

(3) Hun er bager og stolt af det
sub adj

(4) De spurgte her og på bjerget
adv prep + sub

(5) Han lovede bedring og at det ikke skulle gentage sig
sub sætn

The examples show some of the combinations of constituents that may occur in coordinated structures in Danish texts. However, there are certain restrictions with regard to which unlike constituents can be conjoined, but the rules for these restrictions are not easy to determine from a surfaceoriented constituent analysis.

(6) * Hun sang smukt og en arie fra Aida
adv sub

(7) * Han var bager og på bjerget
sub prep + sub

(8) * De spurgte hende og på kontoret
sub prep + sub

The underlying pattern emerges if we look at the syntactic functions of the constituents instead of their grammatical category.

(1): subj v modifier + modifier

(2): subj v attr.subj + attr.subj

(3): subj v attr.subj + attr.subj

(4): subj v modifier + modifier

(5): subj v obj + obj

(6): * subj v modifier + internal.obj

(7): * subj v attr.subj + modifier

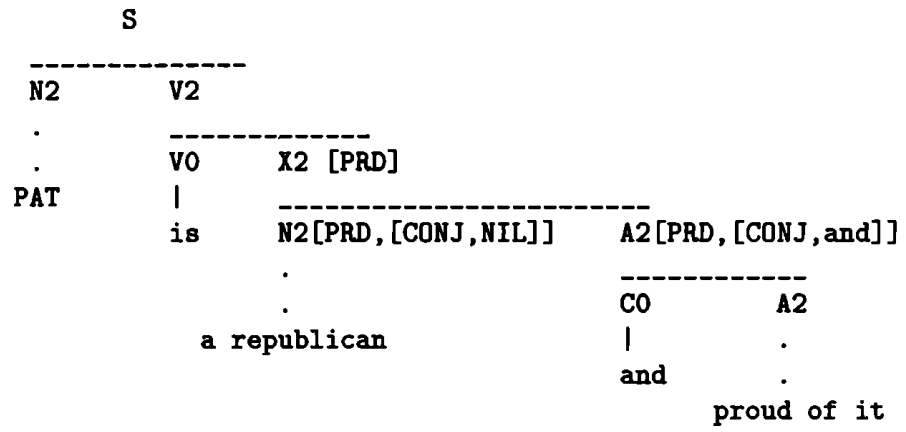
(8): * subj v obj + modifier

Diderichsen (1946) notes that:

> Leddene i en sideordnet Forbindelse staar hvert for sig i samme Forhold til et tredje Led som Helheden

and in this way he introduces the syntactic functions of the constituents as a criterion for a correct coordination. Gazdar et al. (1985) also base their analysis

of coordinated structures on this observation. In GPSG, constituent structure information as well as the syntactic information can be accessed at the same time, and consequently in GPSG the coordination of unlike constituents is treated quite elegantly:



Unfortunately, we cannot just transfer this analysis directly to the Eurotrian model, because of the stratificational design of the system which implies that different information is computed at different levels. This means that we cannot access the information about the syntactic function of the constituents at ECS because it is not computed before the next level. We have to write rules for all possible combinations of constituents at ECS. The result is a provisional overgeneration at ECS before the validation of the constructions can take place at ERS and IS.

4.2 Incomplete Constituents

The most well known account for the coordination of incomplete constituents is probably the generative one advocated by Chomsky in 1965. According to Chomsky, a coordinate surface structure is derived by means of conjunction reduction from two parallel sentences in the deep structure.

Surface structure	(1)	Han elskede huset og haven
=>		
Deep structure	(2)	Han elskede huset
	(3)	Han elskede haven

Simon Dik (1972) led this theory ad absurdum with an example like the following, which he claims leads to 81 different sentences in the deep structure.

Surface structure	(4)	John og Karl og Kurt sælger æbler og pærer og bananer i København, Odense og Århus mandag, tirsdag og onsdag
=>		
Deep structure	4.1.	John sælger æbler i København mandag.
	4.2.	Karl sælger pærer i Odense tirsdag.
	4.3.	Kurt ....
	4.4.	etc. til 81.

The question is now, whether conjunction reduction really is an irrelevant rule or whether the concept or some instance of it could be useful. From the point of view of machine translation one of the first things to be investigated is the translational relevance. Consider the following examples:

(5) I know the woman who painted _ and you met the man who stole the picture that Harry was so fond of _

(5a) jeg kender den kvinde som malede _ og du mødte den mand, som stjal det billede, som harry var så glad for _

(5b) * ich kenne die Frau, die _ malte und du trafst den Mann, der daß Bild stahl, das Harry _ so gern mochte

(5c) ich kenne die Frau, die das Bild, daß Harry so gern mochte, malte, und du trafst den Mann, der es stahl.

In the English and Danish sentences the rules for 'Across-the-board extraction' seem to be the same and we can produce a similar surface structure. In the translation into German (5b) it is obvious that the same mechanism does not work and that the sentence is ungrammatical. In German, different operations have to take place if we want to create an adequate translation. In order to do this, the German generation module must have access to the complete information.

(6) John offered _ and Harry gave Sally a Cadillac

(6a) John tilbød _ og Harry gav Sally en cadillac

(6b) *John bot _ (an) und Harry gab Sally einen Cadillac

In (6)–(6b) the same problem arises, here two constituents are extracted from the first conjunct and again an equivalent surface realization is impossible in German because of the detached preposition "an". (7) shows that this type of extraction is possible if the verb does not have a detached prefix.

(7) John verkaufte und Harry gab Sally einen Cadillac

In (8)–(8b) an equivalent structure cannot be build neither in Danish nor in German.

(8) john put the book away and _ the glass on the table

(8a) *john legte das buch weg und _ das glas auf den tisch

(8b) *john lagde bogen væk og _ glasset på bordet

(8aa) John legte das Buch weg und stellte das Glas auf den Tisch

The verb "put" is translated differently depending on the nature of the object it takes—either to "legen"/"lægge" or to "stellen"/"stille" in German and Danish. A new verb must be introduced in the two target languages to ensure the correct translation of the second part of the clause to match the semantic features of the object.

In these cases a completion of the incomplete constituents by filling the gaps would make the translation much easier.

5 Some Solutions

From the examples in section 4 it is obvious that in some cases gaps have to be filled at IS. We cannot be sure that the incomplete constituents have equivalents in the target language. The process of reduction and extraction is monolingually determined and heavily influenced by phenomena as surface syntax, homonymy and selectional properties of lexical items. This means that we have to leave it up to the target level generator to produce the correct degree of reduction on the basis of the maximal structure at IS. However, there is no need to treat every type of coordination as a reduction. This would lead to ridiculous multiplications of the structures as already noted by Dik.

Therefore, as a working strategy we are pursuing the following approach to these problems:

We distinguish two types of coordinate structures:

1. Coordination of arguments/modifiers
2. Coordination of governors

5.1 Coordination of Arguments/Modifiers

In case of coordination of arguments/modifiers the reduced structure is not rebuilt at IS, but simply kept as the coordination of two or more constituents as in figure 6, which illustrates the following sentence:

(9) John and Peter left

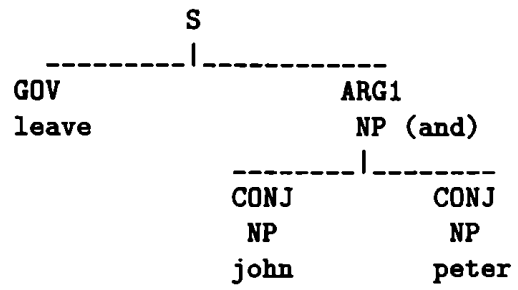


Figure 6:

5.2 Coordination of Governors

In case of coordination of governors (10) the missing arguments are inserted and coindexed with the corresponding constituents in the first conjunct in order to create complete structures as in figure 7, where the coordination of the two verbs is transformed into the coordination of two sentences. Only the valency bound arguments are copied to ensure the completeness and coherence of the structure according to the definition of the IS structure.

(10) We gathered and marched for several hours

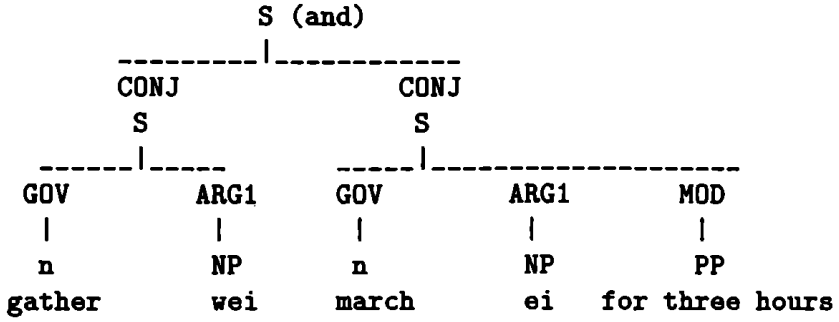


Figure 7:

The completion of the coordinated structure also takes place if more than one gap is found in the coordinated structure as in (8) above, where the second part of the structure consists only of the object (the glas) and the modifier (on the table). In this case both the subject and the verb are inserted and coindexed.

Thus, incomplete structures are only made complete if we are dealing with the coordination of governors or small clauses. We believe that a large amount of problem cases can be solved in this way.

6 Conclusion

Although the theory for basic coordination is well developed and well functioning in the EUROTRA system, there are still a number of problems that we have not mentioned here i.e. the role of negation in coordinate structures, the calculation of features, the determination of the categorial status of a coordinate structure that consists of unlike constituents etc. Some of these have been solved whereas others still are the topic of ongoing research.

For complex coordination the picture is more unclear since we have to deal with gapping both from a monolingual and a translational point of view. However, we believe that the comparative research will give fruitful input to the monolingual analysis.

References

Arnold, D. 1986. General view of the design methodology. *Multilingua* 5–3/1986.

Bresnan, J. [ed]. 1982. *The Mental Representation of Grammatical Relations.* MIT Press, Cambridge, Mass.

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* MIT Press, Cambridge, Mass.

Diderichsen, Paul. 1946. *Elementær Dansk Grammatik.* Gyldendal, København.

Dik, Simon. 1972. *Coordination.* North-Holland, Amsterdam.

Gazdar, G. et. al. 1985. *Generalized Phrase Structure Grammar.* Harvard University Press.

Hansen, A. 1967. *Moderne dansk.* Grafisk Forlag, København.

Jaspaert, L. 1986. The levels of representation. *Multilingua* 5–3/1986.

Jørgensen, P. 1966. *Tysk Grammatik* I–III. Gad, København.

Lang, E. 1984. The Semantics of Coordination, *SLCS* 9. John Benjamins, Amsterdam.

Nirenburg, S. 1989. Knowledge based Machine Translation. *Machine Translation* vol. 3 and 4.

Perschke, S. 1986. Eurotra: General Overview. *Multilingua* 5–3/1986.

Raw, T. et al. 1989. An Introduction to the Eurotra Machine Translation System. *Working Papers in Natural Language Processing* vol. 1, Eurotra—Leuven.

EUROTRA-DK
Københavns Universitet
Njalsgade 80
2300 København S.
Danmark