

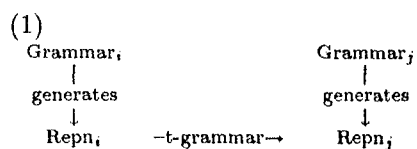
The E-Framework: Emerging Problems

Ian Crookston *
Department of Language & Linguistics
University of Essex
Colchester UK

1 Introduction

Bech & Nygaard (1988) have described a formalism for NLP, the E-Framework (EFW). Two kinds of problem are emerging. Formally, there are problems with a complete formalisation of certain details of the EFW, but these will not be examined in this paper. Substantively, the question arises as to what mileage there is in this formalism for the MT problem. Possibly this question arises about any new NLP formalism, but Raw et al (1988) describe the EFW in an MT context.

The EFW arose in reaction to the CAT formalism for MT (Arnold & des Tombe (1987), Arnold et al (1986)). This was a sequential stratificational formalism in which each level of representation was policed by its own grammar. The essentials of this process can be diagrammed:



*This research has been carried out within the British Group of the EUROTRA project, jointly funded by the Commission of the European Communities and the United Kingdom's Department of Trade and Industry. I am grateful for suggestions and comments from Doug Arnold, Lee Humphreys, Louisa Sadler, Andrew Way, and a COLING reviewer.

The "t-rules" of the t-grammars were the problem.

In the most compact version of this notation, that of Sharp (1988), the t-rules consist of little more than the annotated subtree on each end of the mapping. For instance, in a t-grammar mapping from a predicate-argument representation to a surface one, such as might be necessary in the monolingual modules of an MT system, there might be a t-rule like

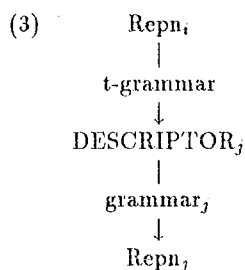
(2)

```
(?,{cat=s}). [ $GOV: (gov,{cat=v}),
               $ARG1: arg1,
               $ARG2: arg2 ]
=>
(s). [ $ARG1,
       vp. [ $GOV,
             $ARG2 ] ]
```

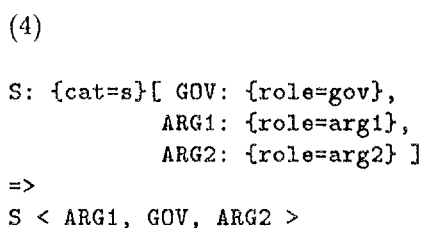
This had the attractiveness of explicitness and clarity, but when it was applied over a wider range of phenomena, two problems emerged. Firstly, the right-hand side (RHS) of the t-rules repeated the target grammar. In (2), the RHS repeats the statement of the surface grammar that the verb and object are dominated by a VP. Secondly, the set of t-rules exploded combinatorially. How this emerged depended on details of the grammars involved. For example, in the above case, if a second

rule were added for passive sentences, then a third and fourth would be needed inserting *will* on the RHS in future active and passive sentences. It is impossible to make separate provision for passive and future—there is a passive simple-tense rule and a passive future rule. To add provision to lower the negation operator into its surface adverbial position now requires not another rule but another four rules.

The general route of the EFW to solving these problems is to separate the output of the t-grammar from the finished representation in the following way, as described in Bech & Nygaard (1988):



The RHS of each t-rule specifies a local part of a special representation called a DESCRIPTOR. This descriptor is then further processed by the grammar to produce a true representation. So (2) becomes



The descriptor formed by the RHS is then, crucially, *parsed* according to the target grammar. All structures defined by the grammar with that mother at the top and those daughters at the bottom are built: specifically, VP is inserted above the verb and object. Given

the extra devices available within grammars, *will* could be added to the descriptor of a future tense sentence after t-rule application, so the four rules covering voice and simple/future tense mentioned above could be reduced to two (or one using the unorderedness device described in Bech & Nygaard (1988)).

This is what makes the EFW the EFW. Without the separate entity called the descriptor being parsed to create the representation, the EFW would lose its defining characteristic. So is there any mileage in this insight for MT?

It should be noted that what is at issue here is the special characteristics of the EFW, those which distinguish it from other stratificational systems such as CAT. The latter have no analogue of the EFW's "consolidation" idea and therefore no analogue of the particular problems discussed here. The EFW is not in these respects representative of stratificational MT formalisms in general.

2 Consolidation and Simple Transfer

What target grammars do to the information specified on the RHS in CAT can be viewed as a special case of what they do in EFW. The EFW can parse, inserting mid-tree nodes and even extra daughters. Successful consolidation includes a judgment that the output representation is acceptable. A special case of this is the geometrical check, where the parser inserts nothing, and delivers an acceptability verdict on the output. CAT is limited to performing the latter special case, which we will refer to as "trivial consolidation". Its complement is "non-trivial consolidation", and it is this latter that is at issue in this section.

There is a well-motivated research programme with the aim of making the transfer

stage of MT as simple as other factors permit (van Eynde (1986); Arnold, des Tombe & Jaspaert (1985, section 2.2.3); Leermakers & Rous (1986)).

Transfer in the EFW is as in (3). The need to make transfer simple must be interpreted as a need to make the two *representations* as similar as possible, rather than to make the source representation and the target descriptor as similar as possible, for methodological reasons. Simple transfer is furthered by researching theories and grammars that assign representations that are similar cross-linguistically. But descriptors are not governed by theories and grammars (other than the bare formal restriction that they must be trees). The only way theories and grammars get access to descriptors in the EFW is by consolidating them into representations. Since there is no theory of descriptors that can be researched in aid of simple transfer, simple transfer must mean similarity of representations rather than of representation and descriptor.

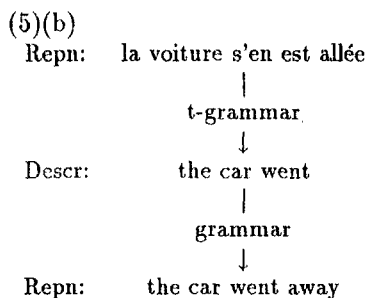
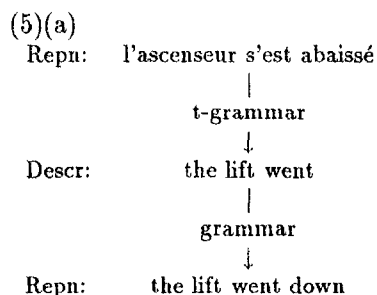
Where the simple transfer research strategy is successful, then source and target representations are identical except for lexical units. Target descriptor and representation must therefore be identical, at least geometrically. Consolidation is therefore trivial checking, as in CAT. Where simple transfer is possible, parsing of descriptors is overkill.

Where simple transfer is not possible, consolidation would be useful if the descriptor resembled the source representation more than the target representation resembled the source.

There are two possible subcases of such a situation. Firstly, there could be a general failure of interface theory to assign similar enough representations. Suppose for example interface theory permitted but did not require VP nodes, and there was a language pair SL and TL such that SL had no VP nodes and TL had them. The TL interface descriptor might then

have no VP nodes, and consolidation could add them to the interface representation. In practice, such language-wide differences in interface geometry seem to be easy to avoid, but in any case there is a problem of modularity. The output of the bilingual component, the transfer t-grammar, is non-trivially altered by a monolingual component, the target interface grammar. This makes the modularity of the bilingual component questionable. In general, it is questionable whether a t-grammar requiring non-trivial consolidation is a separate module from its target generator, and in transfer, this matters, because it is highly desirable that the numerous bilingual components be modular.

Secondly, there can be failures of simple transfer caused by a peculiarity of a specific source expression. If consolidation were used here, this would lead to such processes as:



In such cases, using consolidation to fix up the output of over-simple t-rules is impossible. Such a strategy is always in danger of destroy-

ing the desired translation of some expression other than the one being fixed.

In cases of this latter type t-rules have to take the whole load of the mapping, and the consolidation stage becomes trivial.

So, to recapitulate, where simple transfer is possible, (non-trivial) consolidation is never called upon, and where simple transfer is not possible, it is never useful.

3 Consolidation and Ambiguity

Consolidation has a peculiar property, that in certain restricted circumstances it throws away disambiguation results, recreating earlier ambiguities.

Suppose there is a sentence S that according to grammar G has a set of representations

$$R = \{r_1, r_2, \dots, r_n\}.$$

In the EFW these are trees, but consider them as labelled bracketings. Every r_i will have a "stretch set"

$$S(r_i) = \{r \mid r \in R \ \& \ r \text{ has at least the same brackets as } r_i\}$$

Suppose then that some r_i is consolidated according to a G' that also assigns R as the representations of S . r_i will consolidate ambiguously, yielding $S(r_i)$.

The linguistic claim that this embodies is obviously false. If G and G' are similar enough to yield the same set of representations for S , then each r_i of G is in truth equivalent to a single representation, identical to r_i , in G' : it is not equivalent to a set of G' representations that partly recapitulates the ambiguities that are already identified.

An example is co-ordination. Consider the co-ordinated NP *Bob and Carol and Ted and Alice*. This is 11 ways ambiguous and might be

represented as such at English interface level. The parses include

(6)(a)

```
{np}[ {Bob}, {and}, {Carol}, {and},
      {Ted}, {and}, {Alice} ]
```

(6)(b)

```
{np}[ {np}[ {Bob}, {and}, {Carol} ],
      {and},
      {np}[ {Ted}, {and}, {Alice} ] ]
```

Plausible t-rules into some target language will map these onto descriptors identical to the source representation. Each local tree of these descriptors will then be parsed. It is obvious that this process will in the case of (6)(a) will produce 11 consolidations identical to the original 11 parses. (6)(b) will consolidate unambiguously into something identical to (6)(b). This is because (6)(a) is a case where $S(r_i) = R$, and (6)(b) a case where $S(r_i) = \{r_i\}$. Less flat representations have a smaller stretch set and in this example will consolidate into 3 or 1 translations.

The claim that (6)(a) is 11 ways ambiguous in any target language is false. (6)(a) represents one interpretation of the surface string, and the one interpretation has one translation into any other language.

It is obvious that this weakness also affects the EFW as a formalism for NLP. Suppose (6)(a) were a representation at a predicate-argument level, to be mapped to a surface syntax level of the same language. The EFW embodies a claim that (6)(a) is 11 ways ambiguous on the surface, which is false, just as the claim that (6)(a) has 11 translations is false.

In fact, re-parsing descriptors adds another dimension to the normal problem of the parsing ambiguity of conjoined structures: the first surface parse will in general be ambiguous, and each of its representations will in general map

onto many representations at the next level, each of which will in general breed again at the next level, and so on. Some sample figures are

	Level	No. of Conjuncts			
		2	3	4	5
(7)	1	1	3	11	45
	2	1	5	31	215
	3	1	7	61	595
	4	1	9	101	1269
	5	1	11	151	2321

The number of representations is some function of $l^{(c-2)}$, where l = number of levels and c = number of conjuncts.

4 Consolidation and the Symmetry of Translation

It is often assumed that the relation "possible translation of" is symmetrical (Wanying Jin & Simmons (1986), Isabelle (1988), Arnold & Sadler (1989)). This is plausible: if w in language A translates into x , y , and z in language B, it is surely correct to say that w will appear in the set of possible translations in A of each of x , y and z .

Many MT notations, including CAT, fail to embody this observation in a reversible notation, and thus fail to force linguists to implement a symmetrical translation relation. The EFW makes it impossible to implement a fully symmetrical relation. (6)(a) translates into 11 things in any target language, but (6)(a) does not appear in the set of possible back-translations of any of those 11 except one, the one identical to (6)(a).

5 Conclusion

The descriptor-representation separation and the parsing of descriptors may not be the right

way to tackle the problems of the CAT MT formalism. This is a result which increases the urgency of exploring other avenues to tackling these problems. The obvious other avenue is improving the t-rules themselves, something which is attempted for example in Arnold et al (1988).

6 References

Arnold, D & L des Tombe (1987) "Basic Theory and Methodology in EUROTRA", in S Nirenburg, ed, *Machine Translation: Theoretical and Methodological Issues*, CUP, Cambridge, 114-135

Arnold, D, L des Tombe & L Jaspaert (1985) "Eurotra Linguistic Specifications Version 3", DG XI-II, CEC, Luxembourg

Arnold, D, S Krauwer, L des Tombe & L Sadler (1988), "Relaxed' Compositionality in Machine Translation", in *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Carnegie Mellon Univ, Pittsburgh

Arnold, D, S Krauwer, M Rosner, L des Tombe & G B Varile (1986) "The $\{C, A_i, T\}$ Framework in EUROTRA: A Theoretically Committed Notation for MT", in *Proceedings of the 11th International Conference on Computational Linguistics (COLING 86)*, Association for Computational Linguistics, 297-303

Arnold, D, & L Sadler (1989) "MiMo: Theoretical Aspects of the System", Working Papers in Language Processing 6, Dept of Language & Linguistics, Univ of Essex

Bech, A, & A Nygaard (1988) "The E-Framework: A Formalism for Natural Language Processing", in *Proceedings of the 12th International Conference on Computational Linguistics (COLING 88)*, Association for Computational Linguistics, 36-39

Isabelle, P (1988), "Reversible Logic Grammars for MT", in *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Carnegie Mellon Univ, Pittsburgh

Krauwer, S, & L des Tombe (1984) "Transfer in a Multilingual MT System", in *Proceedings of the*

10th International Conference on Computational Linguistics (COLING 84), Association for Computational Linguistics, 464-467

Leermakers, R, & J Rous (1986) "The Translation Method of ROSETTA", in *Computers and Translation 1*, 169-183

Raw, A, B Vandecapelle, & F van Eynde (1988) "Eurotra: An Overview", in *Interface 3*, 5-32

Sharp, R (1988), "CAT-2—Implementing a Formalism for Multi-Lingual MT", in *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Carnegie Mellon Univ, Pittsburgh

van Eynde (1986) "The interface structure level of representation" in *Multilingua 5*, 145-146

Wanying Jin & R F Simmons (1986) "Symmetric Rules for Translation of English and Chinese", in *Computers and Translation 1*, 153-168