SHAKE-AND-BAKE TRANSLATION

P. Whitelock
Sharp Laboratories of Europe Ltd.
Oxford Science Park
Oxford, U.K.
pete@uk.ac.ox.prg

1. Introduction

In this paper, I describe a view of Machine Translation (MT) that promises radical new solutions to some of the problems of structural approaches such as transfer and interlingua. The new approach is based on a lexicalist view of grammar in which lexical items are multi-dimensional signs. Translation equivalence is defined between bags (multi-sets) of lexical items. Source language analysis determines the bags on one side of the equivalence, and Shake-and-Bake generation combines the corresponding TL signs freely according to the grammar of the target language. The sharing of variables between the logical forms of the signs in the two languages realises the necessary semantic constraints. It is easy to formulate statements of equivalence between expressions having widely divergent syntactic or logical structures, and apply this knowledge decidably for translation in either direction.

2. Structure-Based MT

Perhaps the fundamental question in MT system design is the form in which information about the source text is passed to generation. Such information must include anything relevant for translation, but must be expressed in a form that generation can be guaranteed to make sense of. The answer therefore has important implications for the extent to which the grammars of the languages in the system need be specifically tuned to one another. The ideal is a completely modular approach, with no influence of one monolingual grammar on another – modularity ensures that a system can be easily extended to new languages and language pairs. A satisfactory answer should also provide for reversibility, so that a single modular description of a language may be exploited by both parsing and generation algorithms.

The standard assumption is that all textual information is organised into a sequence of one or more linguistic or logical structures. Transfer-based MT (e.g. Vauquois and Boitet (1985), Nagao et al. (1985), Alshawi et al. (1991), Allegranza et al. (1991) presupposes a language-pair specific module that transforms a structure which is the output of analysis into one that is the input to generation. An interlingual approach (e.g. Uchida and Sugiyama (1980), Lytinen and Shank (1982), Rösner (1986), Nirenburg et al. (1991)) is characterised by the assumption that the output of analysis and the input to generation are isomorphic structures.

The notion of a structure which carries all the information pertinent to translation is common to both transfer and interlingual approaches. In the case of transfer, this is normally a syntactic dependency structure enriched with various other relevant information. Since the syntactic structures of translationally equivalent expressions (TLEs) may differ in many ways, transfer comprises a complex set of operations intended to transform source into target structures. If generation is to be under the control of the same grammatical knowledge as analysis, then the output of transfer from L to L' must correspond to a possible output from the analysis of some expression in L'. Unfortunately, the only way of achieving this is by careful hand-crafting of the transfer grammars. Furthermore, since the invocation of transfer operations is controlled by the structure of the source text, it is problematic to state structural transformations in a declarative, bidirectional manner.

The interlingual approach is seen by its advocates as the solution to this sort of problem. An appropriate system of semantic representation is specified, with the assumption that TLEs will map to identical expressions of such a system. But as Ramsay (1991) points out, only a fraction of the possible sentences of a semantic representation language will correspond directly to natural language expressions. There is ample evidence that the expressions of different natural languages may map to different subsets of the representation language. In particular, TLEs may be associated with distinct semantic representations (cf. the subset problem of Landsbergen, 1987). At minimum, equivalent words may have overlapping or properly subsuming senses rather than identical senses (see e.g Tsujii and Fujita (1991)), but more radical, structural differences are commonplace. Even if TLEs map to logically equivalent expressions, the inference needed to compute one from the other is clearly undecidable in the general case for logics rich enough to represent linguistic meaning (Shieber, 1988), and intractable even in the simplest cases, with associativity and commutativity of conjunction (Calder et al. 1989).

Heuristics to guide such inference can only be specific to the languages in the system, since it is these which determine the logical forms that actually occur. In this way, the functions of transfer merely devolve to analysis and generation, making the content of each grammar dependent on that of the other languages in the system. This weakens the standard argument for the interlingual approach, i.e. that the addition of new languages is facilitated by the absence of components specific to language pairs.

In any approach to translation, it is necessary to state heuristic information concerning the differences in structure between TLEs. In transfer systems this is done explicitly, in interlingual systems it is implicit. I have suggested that in interlingual systems the need for such information has an adverse effect on the independence of different grammars. In multilingual transfer systems, the need for minimal transfer (as advocated within Eurotra (e.g. Allegranza et al., 1991)) also entails that the form of the monolingual components is sensitive to the particular languages in the system (cf. the notion of 'euroversal' in Eurotra). So neither structure-based approach leads to a system in which the knowledge pertinent to the different languages is clearly modularised. The transfer approach is preferable in this respect, but it suffers from the difficulty of formulating declarative statements of complex equivalences, such as those discussed by Kaplan et al. (1989), Odijk (1989), Sadler et al. (1989, 1990), Sadler and Thompson (1991). In §4 below we will look at how such examples are treated in the Shake-and-Bake approach.

3. Shake-and-Bake Translation

The research described in this paper is an application to translation of a more general constraint-based view of language and language processing (e.g. Fenstad et al. (1985), Pereira (1987), etc.). In such a view, linguistic structure is determined by the accumulation of constraints or descriptions, leading to compact and declarative grammars. In Shake-and-Bake MT, we take advantage of the partiality inherent in such constraints by radically underspecifying the information about the source text's structure that is passed to generation.

A precursor to this view can be identified in an approach to MT described by Landsbergen (1987). He suggested that the bilingual knowledge needed to compute TLEs should be stated as correspondances between grammars rather than between languages (that is, the inputs/outputs of grammars). Translation equivalence is stated between the meaningful elements of two grammars – in Landsbergen's Montagovian framework, between lexical entries and between grammar rules. TLEs can be derived from equivalent lexical entries by applying equivalent rules in the same way.

Rosetta's linguistic basis in Montague grammar, with its stipulated pairings of syntactic and semantic rules, requires TLEs to have isomorphic derivations. For this reason, Rosetta is essentially an interlingual system, and the need for the monolingual grammars to be attuned to each other (as acknowledged by Landsbergen) has adverse effects on modularity.

Suppose, however, that the only meaningful elements of a grammar are its lexical items. In fact, much recent linguistic work assumes exactly this conception of a grammar – see e.g Uszkoreit (1986), Karttunen (1989), Pollard and Sag (1987), Zeevat et al. (1987), Baschung et al. (1987), Gunji (1987), Whitelock (1991b). In these frameworks, lexical entries are signs, that is, they simultaneously classify an expression in multiple dimensions (orthographic, syntactic, semantic, etc.), Signs are recursively combined by simple rules which do not themselves introduce elements of meaning, but merely equate appropriate variables in the logical forms of the combining signs.

Therefore in such a grammar, the derivable logical forms in a language are constructed entirely from templates introduced by lexical items. We can represent sentence meaning as a combination of word meanings and a set of equational constraints on LF variables as determined by derivation. Now if we apply Landsbergen's methodology within this framework, knowledge of bilingual equivalence may be reduced to statements of equivalence between lexical items and their associated meanings. Such equivalences may be many to many, e.g. {pay, attention, to} ≡ {faire, attention, à}, and may include multiple instances of the same lexical item, e.g. {as,as} ≡ {aussi, que}[1]. In the general case, therefore, equivalences are stated

[1] The bilingual lexicon is also many-to-many in another sense. A single lexical item in L may appear in many different equivalences with signs in L', and each of the latter may appear in further equivalences with signs in L.

between **bags** of lexical signs. The representation of a sentence is a bag of (extensions of) lexical items, called its **base**. Two bases are equivalent if they are the union of equivalent bags. Two expressions are translation equivalents if they have derivations whose bases are equivalent bags and which obey the same constraints on LF variables. There is no isomorphism requirement on TLEs – the grammars of the two languages have been effectively decoupled.

(1) shows a Prolog definition of a simple translation algorithm based on these principles (| is the path constructor for feature structures, and <=> is graph unification (with path evaluation)).

```
(1) translate(Text,Translation) :-
      SourceSign|orth <=> Text,
      parse(SourceSign,SLBag),
      SourceSign|sem <=> Sem,
      skolemise(Sem),
      equivalence(SLBag,TLBag),
      generate(TLBag,TargetSign),
      TargetSign|orth <=> Translation.
```

translate can be glossed as follows: find the sign in SL whose orthography is the source string **Text**, i.e. parse it to **SourceSign**, recording in **SLBag** the basic expressions (lexical entries) used in the parse. Find the value of semantics in **SourceSign** and Skolemise the variables. The elements of **SLBag** will be extended by their participation in the analysis stage. Alternative extensions may give rise to alternative equivalences. Compute the equivalent bag of TL expressions. Generate any **TargetSign** that can be built from **TLBag** – its orthography is a possible **Translation** of **Text**.

The sharing of those constraints which equate LF variables is realised by Prolog sharing between the variables in **Sem** and the items in **SLBag** and **TLBag**. The Skolemisation step which replaces each distinct variable by a unique constant ensures that variables not equated in the course of analysis are prevented from being incorrectly equated during generation. Its ordering before bilingual lookup realises the restriction that a lexical entry is only applicable when its source language constraints strictly subsume those established by parsing.

A simple backtracking Shake-and-Bake generation algorithm is given in (2). The bag is represented as a Prolog list.

```
(2) generate(Bag,TSign) :-
      shake_and_bake([],TSign,Bag,[]).

% termination
shake_and_bake([Sign],Sign,[],[]).
```

```
% shift
shake_and_bake(P0,Sign,
                     [Next|Bag0],Bag) :-
  push(Next,P0,P),
  shake_and_bake(P,Sign,Bag0,Bag).

% reduce
shake_and_bake(P0,Sign,Bag0,Bag) :-
  pop(First,P0,P1),
  delete(Second,P1,P2),
  unordered_rule(Mom,First,Second),
  push(Mom,P2,P),
  shake_and_bake(P,Sign,Bag0,Bag).
```

shake_and_bake is a slight but significant variation on a shift-reduce parser for binary grammars. In shift-reduce parsing, an element is repeatedly taken from the front of the input, looked up in the lexicon, and the result pushed onto a stack. The top two stack elements may be combined according to the grammar, the result pushed back on the stack, and the procedure called recursively. When the input has been consumed and the stack contains a single element, the parse terminates successfully.

In **shake_and_bake** generation the role of the input string is played by the bag of target language signs. Unlike in parsing, the order of items in the bag is immaterial. Any two signs may be combined, with the combination determining the order of their orthographies in the result. A minimal complete algorithm requires random access to the erstwhile stack for one of the arguments to a reduction. This data structure is represented by the variables of the form P0, P1, P. The second element is not popped from the data structure, but non-deterministically deleted from it. Of course, such a move renders the algorithm intractable. Shift-reduce can complete a well-formed symbol table or chart in polynomial time for context-free grammars, but Shake-and-Bake is exponential even with a chart. I will mention ways of addressing this computational drawback below. Before this, however, I will try to show that an MT system organised in this way makes it considerably easier to correctly state constraints on translation equivalence when TLEs have divergent structures. Furthermore, such statements of equivalence may be exploited in either direction.

4. Translation Equivalence in Shake-and-Bake MT

I will use the PATR-II notation for equations (Shieber 1986), representing constraints on the elements of an equivalence rather than the categories in a grammar rule. The substantive theory could be any of the lexicalist grammars mentioned above. I assume a morpheme-based lexicon in which each lexical entry (i.e. morpheme) has a feature **cite** whose value uniquely picks out that entry. This feature is used to distinguish words spelled the same but with different syntax or semantics. In addition, where the orthographic form of a grammatical morpheme (such as past tense) is determined on a language internal basis (e.g. by conjugation type of a verb), all allomorphs will receive the same value of **cite**. In this way, the notion of equivalence may be extended to equivalence between closed-class grammatical morphemes.

(3) shows a basic one-to-one equivalence between the English verb stem *love* and the French *aimer*.

(3) $X_E \equiv X_F$

<X_E cite> = love_v
<X_F cite> = aimer
<X_E sem index> = <X_F sem index>
<X_E sem exp index> =
        <X_F sem exp index>
<X_E sem obj index> =
        <X_F sem obj index>

I assume an event-based semantics with a neo-Davidsonian treatment of thematic relations (see e.g. Parsons (1980, 1991), Hobbs (1986), Dowty (1988)). The two monolingual signs presupposed by (3) each introduce three LF variables, corresponding to the loving state itself (**index**), the lover (**exp index**) and the loved one (**obj index**). (3) states the simple pairwise equivalence between these. Despite the identical paths on both sides of the last two equations, the correctness of the method does not rely on thematic identity of equivalent roles. It may apply equally to grammars where thematic relations are verb-specific (e.g. love_arg1, love_arg2, aimer_arg1 etc.) and hence language-specific (sidestepping consistency problems in the monolingual assignment of thematic relations). Furthermore, even with a substantive theory of thematic relations applied consistently to two grammars, the thematic entailments of equivalent argument places may not be identical. Space precludes discussion of our treatment for such cases, which is based on the use of sorted LF variables as described in Moens et al. (1989).

Assuming equivalences such as (4) for proper names, translation between (5a) and (5b) will be mediated by the equivalence between bags shown in (6), in which signs are abbreviated by their citation values.

(4) $X_E \equiv X_F$

<X_E cite> = Mary
<X_F cite> = Marie
<X_E sem index> = <X_F sem index>

(5a) Mary loves Frances
(5b) Marie aime Françoise

(6) {Mary, Frances, love_v, pres} ≡
    {Françoise, pres, aimer, Marie}

Generation will produce (5a) as the translation of (5b), and vice versa, not because that was the structure of the SL text, but because it is the only way of putting together the TL signs in (6) that is compatible with the TL grammar and the variable bindings established by parsing.

This strategy extends straightforwardly to the 'argument switching' cases of translation equivalence exemplified in (7a and 7b).

(7a) Mary likes Frances
(7b) Françoise plaît à Marie

The equivalence between *like* and *plaire à* is given in (8).

(8) $X_E \equiv X_F$ & X_F'

<X_E cite> = like_v
<X_F cite> = plaire
<X_F' cite> = à_1
<X_E sem index> = <X_F sem index>
<X_E sem exp index> =
        <X_F sem exp index>
<X_E sem obj index> =
        <X_F sem obj index>
<X_E sem exp index> = <X_F' sem index>

I have assumed that the path <X_F sem obj> picks out the semantic object of the liking state, which is the semantics of *plaire*'s syntactic subject, but as in the previous example, alternative monolingual assumptions could have been made with little import.

One point to note here is the treatment of the preposition *à*. We merely state that one lexical entry in English corresponds to two in French. The appropriate preposition (<cite> = à_1), is an element with 'identity semantics' (Calder et al. 1989), so its index is equated in the monolingual lexicon with that of its syntactic object. It is also necessary to equate this index with the experiencer of *like* and *plaire* as in the final equation. Then the strict subsumption interpretation of bilingual lookup will guarantee that the elements on the French side are not just unrelated elements of a larger phrase.

Examples of the argument-switching kind are standard in the transfer approach as illustrations of what it does best. Since a typical transfer MT system will process a functional structure by recursive descent, cases where the translation of arguments depends on the translation of heads are straightforward. However, a major source of problems for such transfer schemes are the so-called head-switching cases, in which the syntactic head-dependent structure in the translation equivalents is reversed (see refs. in §2). Such a case is illustrated for English/Dutch in (10). The *just /venir de* case in English/French is similar.

```
(10a) Jan zwemt graag
(10b) John enjoys swimming
```

Once again, the Shake-and-Bake generation strategy will correctly compute such TLEs from statements of equivalence between simplex expressions, including that given in (11). The full bags are given in (12).

(11) X_E & X_E' ≡ X_N

<X_E cite> = enjoy
<X_E' cite> = prespart
<X_N cite> = graag
<X_E sem index> = <X_N sem index>
<X_E sem exp index> =
<X_N sem exp index>
<X_E sem obj index> =
<X_N sem obj index>
<X_E' sem index> = <X_N sem index>

```
(12) {jan,pres,zwemen,graag} ≡
       {john,pres,enjoy,prespart,swim}
```

The fact that equivalent tense morphemes (*pres*) occur on non-equivalent stems (*enjoy /zwem+*) follows immediately from the mechanics of generation. Whitelock (1991a) includes further discussion of these and other examples, which include a further head switching case in English/French/Japanese equivalence (13a,b,c), and an example of distinct but equivalent logical forms in English/Japanese (14a,b)

```
(13a) John runs up the street

(13b) Jean monte la rue en courant

(13c) Zyon wa hasite toori   wo  noboru
   John TOP running street ACC goes up

(14a) (p -> q) Maria eats only fish

(14b) (~q -> ~p)
    Maria wa sakana shika tabe-na-i
   Maria TOP fish     ?   eat-NEG-PRES
```

5. Conclusions and Further Research

A small trilingual (English/French/Japanese) system based on the above ideas has been implemented and is currently under development. Beaven (1991) describes a similar implementation of an English/Spanish system. The modularity of the monolingual grammars remains uncompromised even if TLEs have radically different syntactic and/or logical structures, since all contrastive knowledge is clearly localised in the bilingual lexicon. Although this paper has only touched on the question of complex equivalences, it is clear from our intial implementations that the declarative description of these is massively simplified by the adoption of a Shake-and-Bake approach. The bilingual lexicographer is not required to specify equivalences between abstract structures at a single (hybrid) level of representation, as in a transfer system. Instead, equivalence is stated between bags of multi-dimensional lexical signs. Constraints on any dimension of such signs may be included (cf. Tsujii, 1986 for the necessity of such multi-dimensional constraints). We believe that the concrete nature of such a task offers interesting possibilities for the automated acquisition of bilingual correspondences from aligned corpora. We therefore also see Shake-and-Bake as a first step in the integration of traditional linguistic (rationalist) and novel statistical (empiricist) approaches to MT (e.g. Brown et al. (1990)).

Since Shake-and-Bake generation is an NP-complete problem (see Brew, this conference), there is no tractable general algorithm. In order to improve average case performance, we need to be able to impose further structure on the bag which forms the input to generation. For example, the syntactic structure of the source text might be called upon to provide heuristic control information for generation. The resulting system would be similar in character to a transfer organisation, but avoiding the 'procedural seduction' of Kaplan (1987), that is, the temptation to allow or require a grammar writer to state detailed control information. While it seems that such a possibility would enable a grammar to be applied more efficiently, Kaplan argues that a computer is almost always better qualified to carry out the task of optimising the procedural interpretation of a large coverage grammar. In the current context, the procedural seduction is that the bilingual grammar writer is the one best qualified to define the structural 'changes' contingent on the definition of particular lexical equivalences. One approach to the efficiency problem will therefore address the automatic determination of useful control information from the derivation of the source bag and the exploitation of this in generation. Brew (this conference) presents an alternative approach which extends van Benthem's Count Invariant for Categorial Grammars (see e.g. Moortgat, (1988)) to constraint grammars. In this way, fruitless branches of the search space may be pruned early, thus improving efficiency of generation using monolingual TL information.

In the discussion so far, it has been assumed that the only functions of syntax in parsing have been a) to extend the lexical signs and thereby constrain the applicable equivalences, and b) to determine the bindings of LF variables and thus provide the semantic constraints that are the invariants for translation. We have not considered a class of invariants that may be grouped under the heading of discourse structure. In fact, logically equivalent discourse variants are typically associated with non-equivalent bags; for example, the passive morpheme may be present in one but not another; word order features may also be constrained in a bilingual equivalence. So such invariants can be captured. On the other hand, a more elegant treatment might involve a computation of a set of interlingual discourse constraints on derivations to be used in generation in the same way as the semantic constraints on LF variables. Clearly such an approach would be very similar to an interlingual system, but without the adverse consequences for grammatical modularity characteristic of such systems.

Acknowledgements

The ideas and algorithms presented here were developed jointly with Mike Reape. John Beaven and Evelyn van de Veen in Edinburgh helped by exploring them in their PhD and MSc theses respectively. Sharp Laboratories of Europe gave me the opportunity to continue working on them, and I would like to acknowledge the help of all my colleagues there, especially Ian Johnson, Chris Brew, Olivier Laurens and Tosh Nomura.

References

Allegranza, V. , P. Bennett, J. Durand, F. van Eynde, L. Humphries, P. Schmidt and E. Steiner (1991) "Linguistics for Machine Translation: The Eurotra Linguistic Specifications". *Studies in MT and NLP*, CEC.

Alshawi, H., D. M. Carter, B. Gambaeck and M. Rayner (1991) "Interactive Translation using Quasi-Logical Forms", *Proceedings of the 29th ACL*, Berkely.

Baschung, K., G. Bes, A, Corluy and T. Guillotin (1987) "Auxiliaries and Clitics in French UCG Grammar", *Proceedings of the Third European ACL*, Copenhagen.

Beaven, J.L. (1991) *Lexicalist Unification-Based Machine Translation*, PhD Thesis, submitted to Dept. of Artificial Intelligence, Univ. of Edinburgh.

Brew, C.H. (1992) "Letting the cat out of the bag: Generation for Shake-and-Bake MT", this conference.

Brown, P., J. Cocke, S. della Pietra, V.J. della Pietra , F. Jelinek, J.D.Lafferty, R.L. Mercer and P. S. Roossin (1990) "A Statistical Approach to Machine Translation", CL vol. 16, no 2.

Calder, J. , M. Reape & H. Zeevat (1989) "An Algorithm for Generation in UCG", *Proceedings of the Fourth European ACL*, Manchester.

Dowty, D. (1989) "On the Semantic Content of the Notion 'Thematic-Role'", in G. Chierchia, B. Partee and R. Turner (eds.) Property Theory, *Type Theory and Natural Language Semantics*, Dordrecht, Reidel.

Fenstad, J.E., P.K. Halvorsen, T. Langholm and J. van Bentham (1985) *Equations, Schemata and Situations: A Framework for Linguistic Semantics*. CSLI-85-29, Stanford.

Gunji, T. (1987) *Japanese Phrase Structure Grammar*, Reidel, Dordrecht.

Hobbs, J. (1986) "Ontological Promiscuity", *Proceedings of the 25th ACL*.

Kaplan, R. M. (1987) "Three Seductions in Computational Psycholinguistics", in P.J. Whitelock, M.M. Wood, H. Somers, P. Bennett, R. Johnson (eds.) *Linguistic Theory and Computer Applications*: Academic Press.

Kaplan, R. M., K. Netter, J. Wedekind and A. Zaenen (1989) "Translation by Structural Correspondances", *Proceedings of the Fourth European ACL*, Manchester.

Karttunen, Lauri (1989) "Radical Lexicalism", in M.R. Baltin and A.S. Kroch (eds.), *Alternative Conceptions of Phrase Structure*, Chicago.

Landsbergen, J. (1987) "Montague Grammar and Machine Translation", in P.J. Whitelock, M.M. Wood, H. Somers, P. Bennett, R. Johnson (eds.) *Linguistic Theory and Computer Applications*:

Academic Press.

Lytinen, S and R. Shank (1982) "Representation and Translation". TR 324, Dept. of Computer Science, Yale University.

Moens, M., J. Calder, E. Klein, M. Reape and H. Zeevat (1989) "Expressing Generalisations in Unification-Based Grammar Formalisms", *Proceedings of the Fourth European ACL*, Manchester.

Moortgat, M. (1988) *Categorial Investigations: Logical and Linguistic Aspects of the Lambek Calculus.* Foris, Dordrecht.

Nagao, M. J. Tsujii and J. Nakamura (1985) "The Japanese Government Project for Machine Translation", *Computational Linguistics*, vol. 11 #2-3

Odijk, J. (1989) "The Organisation of the Rosetta Grammars", *Proceedings of the 4th European ACL*, Manchester.

Parsons, T. (1980) "Modifiers and Quantifiers in Natural Language", *Canadian Journal of Philosophy*, supp. vol. VI.

Parsons, T. (1991) *Events in the Semantics of English*, MIT Press, Cambridge, Mass.

Pereira, F. C.N. (1987) "Grammars and Logics of Partial Information", in *Proceedings of the 4th International Conference on Logic Programming*, Melbourne, Australia.

Pollard, C and I. Sag (1987) *Information-based Syntax and Semantics: Vol 1: Fundamentals*, CSLI Lecture Notes 13.

Pollard, C and I. Sag (forthcoming) *Information-based Syntax and Semantics: Vol 2*, CSLI.

Ramsay, A (1991) "A common framework for analysis and generation", in *Proceedings of the 5th European ACL*, Berlin.

Rösner, D. (1986) "When Mariko talks to Siegfried: Experiences from a Japanese/German MT project". *Proceedings of the 11th International Conference in Computational Linguistics* (COLING), Bonn.

Sadler, L., I. Crookston and A. Way (1989) "Co-description, projection and 'difficult' translation", *Working Papers in Language Processing* #8, Dept. of Language and Linguistics, University of Essex.

Sadler, L., I. Crookston, D. Arnold and A. Way (1990) "LFG and Translation", in *Third International Conference on Theoretical and Methodological Issues in MT*, Linguistics Research Centre, Austin, Texas.

Sadler, L and H. S. Thompson (1991) "Structural Non-Correspondance in Translation", in *Proceedings of the 5th European ACL*, Berlin.

Sheiber, S.M. (1986) *An Introduction to Unification-Based Approaches to Grammar*, University of Chicago Press.

Shieber, S. M. (1988) A Uniform Architecture for Parsing and Generation", *Proceedings of the 12th International Conference in Computational Linguistics* (COLING), Budapest.

Tsujii, J. (1986) "Future Directions of Machine Translation", *Proceedings of the 11th International Conference in Computational Linguistics* (COLING), Bonn.

Tsujii, J and K. Fujita (1991) "Lexical Transfer based on Bilingual Signs", *Proceedings of the 5th European ACL*, Berlin, April 1991.

Ushida, H and K. Sugiyama (1980) "A Machine Translation system from Japanese into English based on Conceptual Structure". *Proceedings of the 8th International Conference in Computational Linguistics* (COLING), Tokyo.

Uszkoreit, Hans (1986) "Categorial Unification Grammars", *Proceedings of the 11th International Conference in Computational Linguistics* (COLING), Bonn.

van de Veen, E. (1990) *Parsing Free Word Order Languages*, MSc Thesis, Dept. of Artificial Intelligence, University of Edinburgh.

Vauquois, B. and Ch. Boitet (1985) "Automated Translation at Grenoble University", *Computational Linguistics*, vol. 11 #1.

Whitelock, P. (1991a) "Shake-and-Bake Translation". in *Proceedings of the Workshop on Constraint Propagation and Linguistic Description*. ed. C.J. Rupp, M. Rosner and R. Johnson, IDSIA, Lugano.

Whitelock, P. (1991b) *A Lexicalist Unification Grammar of Japanese.* PhD Thesis, submitted to Dept. of Language and Linguistics, UMIST.

Zeevat, H., E. Klein and J. Calder (1987) "An Introduction to Unification Categorial Grammar", in N.J. Haddock, E. Klein and G. Morrill (eds.) *Edinburgh Working Papers in Cognitive Science, vol. 1: Categorial Grammar, Unification Grammar and Parsing*. Centre for Cog. Science, Edinburgh.

SHAKE-AND-BAKE TRANSLATION

P. Whitelock
Sharp Laboratories of Europe Ltd.
Oxford Science Park
Oxford, U.K.
pete@uk.ac.ox.prg

Résumé

Dans cet article, je décris une conception de la Traduction Automatique qui apporte des solutions nouvelles et radicales à quelques uns des problèmes rencontrés par les approches structurales telles que les modèles à transfert ou à langue pivot. Cette nouvelle approche est basée sur une conception lexicaliste de la grammaire où les unités lexicales sont des signes multi-dimensionnels. La relation de traduction est définie comme une équivalence entre des "bags" (multi-ensembles) d'unités lexicales. L'analyse de la Langue Source détermine les "bags" d'un côté de l'équivalence et la génération par l'algorithme Shake and Bake combine librement les signes correspondants en fonction de la grammaire de la Langue Cible. Le partage des variables entre la forme logique des signes dans les deux langues fournit les contraintes sémantiques nécessaires. Il est facile de formuler des équivalences entre des expressions ayant des structures syntaxiques ou logiques largement divergentes et d'utiliser ces connaissances de manière décidable pour effectuer des traduction dans l'une ou l'autre direction.

要旨

本論文の目的は、変換手法・中間言語手法等構造的アプローチで生じる諸問題に対する新たな解決策を提供する機械翻訳の新観点を述べることである。このアプローチでは文法の語彙的観点を基本としており、語彙項目は多次元記号として取り扱われる。翻訳対応関係は語彙項目の袋（マルチ集合）の対で定義される。ソース言語解析により対応関係の一方が決定され、更に、シェークアンドベーク生成法により、ターゲット言語の文法に従った対応するターゲット言語記号が結合される。また、両言語における記号の論理形式間の変数共有化により、必要な意味制約を実現することができる。この新方式により、全く異なった構文・論理構造をもつ二表現の対応関係を容易に記述することができ、また、この知識構造を双方向翻訳方式に適用することが可能となる。