Machine Translation for Monolinguals

Mary McGee WOOD
Department of Computer Science
University of Manchester
Manchester M13 9PL U.K.

Brian J. CHANDLER
Centre for Computational Linguistics
UMIST
Manchester M60 1QD U.K.
and
International Computers Limited

Abstract

We describe sister machine translation prototypes, Ntran, an English to Japanese system developed at UMIST, and Aidtrans, Japanese to English, at Sheffield, both designed for use by an English monolingual. Aidtrans uses extensive and sophisticated collocational analysis radically to reduce the need for conventional post-editing. Ntran offers interactive query at three stages: on-line dictionary update, syntactic disambiguation, and Japanese lexical selection. The second of these is described and illustrated in particular detail, and the underlying philosophy of monolingual interaction discussed.

1. Background

Under the Alvey Directorate's research programme in natural language processing, an English-Japanese machine translation project was carried out at the Centre for Computational Linguistics, University of Manchester Institute of Science and Technology and the Centre for Japanese Studies, University of Sheffield. The project ran from October 1984 to December 1987, and was also funded by International Computers Limited (ICL). The UMIST group, led by Peter Whitelock, have developed an English-to-Japanese prototype (Ntran), while Jiri Jelinek's group at Sheffield have been working from Japanese into English (Aidtrans).

The two prototypes, although very different in some aspects of their linguistic and computational approaches, share an important and distinctive design philosophy. Both are interactive, and, unlike present commercially available machine (aided) translation systems, both are designed for end use by a monolingual speaker of English. This paper will discuss the means by which each system achieves this, and the issues involved in tailoring a system for use by a target language or source language speaker.

2. Aidtrans: the Sheffield Japanese-to-English system

The Aidtrans Japanese-to-English prototype (implemented in C, and running on a Sharp Unix-based microcomputer) is an implementation of a comprehensive, highly detailed and sophisticated algorithmic grammar of Japanese developed by Dr. Jiri Jelinek as a teaching tool for rapid intensive instruction in technical Japanese (Jelinek 1978). The core of this grammar is its Integrated Dictionary System (IDS). The philosophy of IDS is to incorporate as much as possible of the grammar and the analysis heuristics in the dictionary. This is done in an explicitly language-specific, and, as applied to translation, language-pair specific form, allowing great accuracy and precision (at some inevitable cost in adaptability). The dictionary of the finished prototype contains entries for some 6,000 words.

While committed to the maximum use of lexical resources, Aidtrans also sees translation as a relation over whole texts rather than individual words or even sentences. The purpose of each act of translation is to retain the global sense, rather than the concatenation of word-meanings, of a text as it is reformulated in a different language. To achieve this, it is clearly not enough to produce one acceptable translation for each separate sentence of a text and adjoin them. Just as a syntactic parser will produce alternative analyses of an ambiguous sentence from which the one intended must be selected, so Aidtrans produces alternative translations of each part of the input text, from which the translation most appropriate to the context must be selected.

Such selection from among possible translation equivalents is familiar from human translation or post-editing. Here, however, much of the selection, or rejection, is carried out by the system itself. A text-type-specific linear predictive model is the basis for determining priorities or preferences among the possibilities. Patterns can be recognized at the general level of syntactic configuration and at the more specific level of individual lexical items and collocations; at present the system recognizes well over 200 different types of juxtapositional linkage. In other words, the selectional function in Aidtrans is driven by a generalization of valency, augmented with priority weightings for the possible valency values.

3. Ntran: the UMIST English-to-Japanese system

Ntran - its design inspired by Rod Johnson, and developed and first implemented largely by Peter Whitelock - is less target-specific than Aidtrans. The prototype is implemented in Prolog for the sake of rapid and perspicuous development; versions now exist in Cprolog, New Improved (Edinburgh) Prolog, and Quintus. During the course of development, versions have been run on a DEC MicroVax II, an ICL PERQ, and most recently a Sun 3/50 workstation.

Through a system of nested menus, Ntran functions on three levels: as a system development system, a grammar development system, and a translation system proper. Each level offers specific facilities for the writing, testing and debugging of appropriate areas of program code. (For details, see Whitelock et al 1986).

Although both prototypes give the maximum weight and information content to the lexicon, another point of difference between them is that Ntran is committed to the principle of translation as linguistics (cf. Johnson 1987), and designed and implemented as an explicit embodiment of contemporary lexicalist linguistic theory. The English analysis grammar is based on Lexical-Functional Grammar (Bresnan, ed. 1982) and Generalized Phrase Structure Grammar (Gazdar et al 1985), the Japanese generation grammar on Categorial Grammar (Steedman 1985, Whitelock 1987).

In analysis, words are first looked up in an English morpho-syntactic dictionary which specifies grammatical category and morphologically determined features such as tense and number. The entries in this, as in all dictionaries, are compacted by "feature co-occurrence restrictions" which factor out any feature-values which are predictable on the basis of others. These derive largely from the fcrs of Generalized Phrase Structure Grammar (Gazdar et al 1985). In English, for example, any lexical item which has tense must be finite and a verb. In a lexical entry assigning any value to "tense", the specification of finiteness and verb-hood would be redundant, and can be supplied by a generalized rule of the form

```
fcr(tense=_,[fin=finite,stemtyp=verb]).
```

Similarly, as any verb has no noun features, but sets (possibly empty) of prepositional complements and adjuncts, and as any '-ing' form is a progressive finite verb, we have rules

```
fcr(cat=verb,[nounfeats=[],pcomp=set(_),adjunct=set(_)]).
fcr(nfform=ing,[stemtyp=verb,aspect=progress,inf=no]).
```

Using this limited information, the parser builds all possible "functional structures" (the "f-structures" of LFG), which serve as an intermediate representation abstracting away from surface constituent structure, a particularly valuable level when mediating between a configurational language such as English and a non-configurational one such as Japanese.

A second stage of lookup in the English "subcat" dictionary, which holds possible subcategorization patterns, eliminates spurious f-structures, and provides a semantic interpretation ("s-structure") for those which remain. (Cf Wood et al 1987.) S-structure forms the basis for transfer, driven by bilingual dictionaries, the only component to hold contrastive information. The resulting Japanese s-structure is the basis for generation of a Japanese f-structure, using syntactic information held in the bilingual and Japanese dictionaries in the form of the complex categories and combination rules of a unification categorial grammar (see Whitelock 1988 for details). Surface ordering of the Japanese output is finally carried out by linear precedence rules. The role and form of user interaction will be discussed below.

4. Techniques for interactive translation

As mentioned earlier, both Aidtrans and Ntran are designed for an English monolingual end-user. This approach - reflected in the joint project's Alvey title, "Read and write Japanese without knowing it" - distinguishes them from currently commercially available machine (aided) translation systems, and has led to a number of distinctive design decisions.

4.1 Aidtrans

In the case of Aidtrans, the intention was, while leaving the final selection of the exact translation to the end-user, to produce output of greater accuracy and coherence than is generally found in current post-editing systems. The strategy of multiple generation produces a set of complete alternative translations, rather than one which must be amended piecemeal by a posteditor, while the text-type-based predictive model and preference-weighted linkages cut down greatly on the range actually offered to the end-user, and group those which survive into semantically and stylistically coherent wholes. Thus, while a conventional posteditor needs access to the source text to check the accuracy of raw output and as a guide to its revision, here enough information is available in the output to form the basis of the end-user's final selection.

4.2 Ntran

The facilities for, or demands on, the end-user of Ntran are somewhat more complex: both the complexity of the task and the inner articulation of the system are greater, giving both the need and the opportunity for a variety of interactions (cf Johnson & Whitelock 1987). To ensure to an English monolingual technical writer the output of accurate and acceptable Japanese, the conventional strategy would be pre-editing, passing to the machine only text in a restricted sub-language known to be within its translation capacity. Our system could perhaps be said to offer interactive pre-editing interleaved with translation, rather than interactive translation proper, as no contrastive or bilingual information is presented to the end user in the interaction. The restricted input sub-language, however, is simply grammatical English, which if ambiguous must be disambiguated. This should be seen not as a constraint on a technical writer but as a desideratum.

The Ntran prototype is designed to offer three forms of interactive query: on-line dictionary creation, syntactic disambiguation of English input, and Japanese lexical selection in transfer.

4.2.1 Dictionary update

When a word is found in an input text for which no dictionary entry yet exists, the user is offered the option of creating an entry for it immediately. This is done using a tree-structured question procedure, eliciting the category of the English word and its values for the features associated with that category, such as mass/count and animacy for nouns, valency and aspectual type for verbs, gradability for adjectives, and so on. The on-line dictionary building routine, although it incorporates a reasonable range of information about an English word, does not ask for Japanese translation equivalents. Instead, entries created in this way are held in a separate dictionary file, where they are accessible to the analysis component, but also set aside for later completion by a bilingual linguist.

Until this is done the English word is at present simply passed into the Japanese output in its original form, marked off by a special delimiting character. We intend to implement in a further developed version of the system a facility for passing through such words in katakana transcription. Given a reasonable core dictionary, most new words will be specialized technical terms, for which this will in fact be the correct rendering.

4.2.2 Syntactic disambiguation

Syntactic ambiguities in the English input are also referred to the user for disambiguation. The parser first builds a surface syntactic dependency structure, or functional structure, which is then mapped to a deep or semantic structure, and a record kept of the mappings ('obj', for example, is mapped to 'arg0'). During this mapping stage, a record is also kept, for each well-formed s-structure produced, of the set of mappings entailed by the subcategorization requirements of the lexical items involved. Each mapping records the derived semantic relation which is assigned between a constituent and its parent. Examples of "maptrace" are given with the examples below.

The disambiguation module then computes a set of differences among all the recorded mapping sets and builds a set of all those relations which are true for only a subset of the parses. These are then presented to the user, after conversion of some of the internal semantic relation names to external names which are intended to be more immediately understandable. The generator for the user-form representation of mappings is:

```
describeas(map(X,Y,Z),[X,' is ',C,' of ',Z]) :-
        logtocase(Y,C),!.

logtocase(arg0,object).
logtocase(arg1,agent).
logtocase(ben,beneficiary).
logtocase(loc,location).
logtocase(rep,representation).
logtocase(instr,instrument).
logtocase(adjunct,modifier).
logtocase(X,X).
```

It should be noted that this mechanism succesfully represents both purely structural ambiguities such as prepositional phrase attachment, and also subcategorization ambiguities, as in "write on the deck of the ship", where "deck" could be either the location or object of "write".

The alternatives are presented as a set of statements distinctively characterizing the possible semantic interpretations, as can be seen in the examples below. The user responds with the number of any statement which is true, or "f" followed by the number of any statement which is false.

Because of a technical implementation detail, constituents are at present referred to only by their heads: thus, in this example set of queries, "and is object of active" means "(workstations and terminals) is object of active". Obviously this aspect of the presentation could be improved in a more fully developed system. One could also present the alternatives in quite different ways, by paraphrases of alternative readings, for example, or dependency trees or some other graphical interface, generated by the same underlying mechanism.

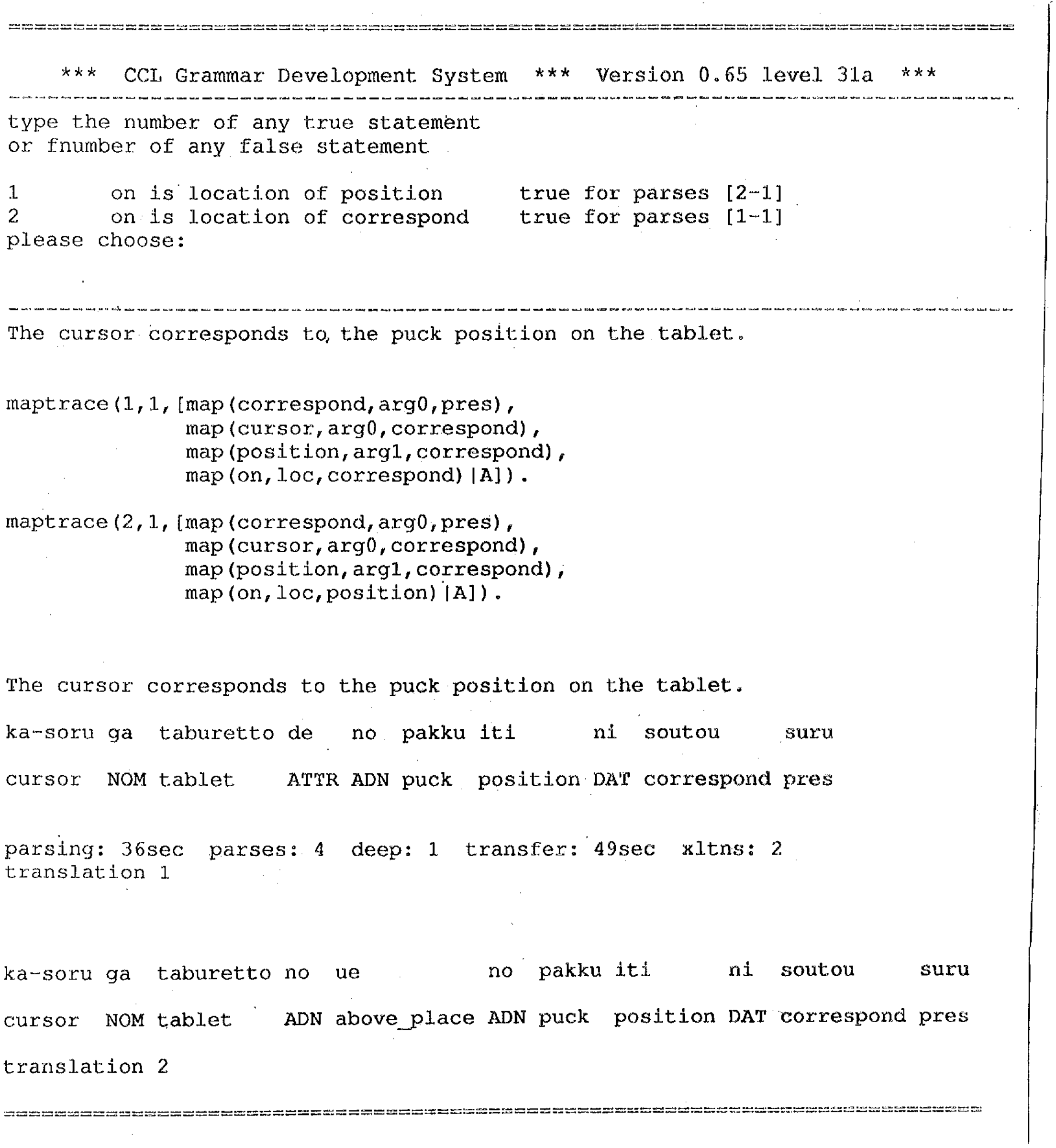
4.2.3 Japanese lexical selection

Finally, ambiguities, or alternatives, may arise in the selection of a Japanese translation equivalent for an English word or expression. Interactive systems standardly offer such alternatives directly to the user, who must have some competence in the target language to be able to make the choice. Ntran's Japanese dictionary entries include English glosses, and the user will be offered these to chose between, rather than the Japanese head-words. This facility is not yet fully implemented.

5. The system as translator and the monolingual user

Clearly, ensuring reliable translation for a monolingual user in either direction requires a system design carefully tuned to the task. In the case of "import translation", translating into the user's language, the information content of the output text must be sufficiently rich that, in cases of uncertainty, reference to the source text (the traditional recourse of the post-editor) is adequately replaced by reference to the set of coherent possibilities offered in that output. This is exactly the strategy implemented in Aidtrans.

In the case of "export translation", when the user is a speaker of the source language, the system can request additional information at a number of stages in the translation chain, to supplement the information inherent in the surface form of the input text, if that proves insufficient for syntactic analysis, semantic interpretation, and/or target language lexical selection. (Although the obvious, and ultimate, source of such supplememtary information is the human end-user, we envisage the long-term possibility of referring queries first to intelligent, world-knowledge-based modules within the system, leaving the human user as a progressively less often needed safety net.) Ntran's modularity of design isolates the stages of the process clearly from each other, while our commitment to the implementation of linguistic theory offers formats for the presentation of choices by the system and the input of information by the user which are transparent to both.



```
    ***  CCL Grammar Development System  ***  Version 0.65 level 31a  ***
--------------------------------------------------------------------------------
type the number of any true statement
or fnumber of any false statement

1       on is location of position     true for parses [2-1]
2       on is location of correspond   true for parses [1-1]
please choose:

--------------------------------------------------------------------------------
The cursor corresponds to, the puck position on the tablet.


maptrace(1,1,[map(correspond,arg0,pres),
              map(cursor,arg0,correspond),
              map(position,arg1,correspond),
              map(on,loc,correspond)|A]).

maptrace(2,1,[map(correspond,arg0,pres),
              map(cursor,arg0,correspond),
              map(position,arg1,correspond),
              map(on,loc,position)|A]).


The cursor corresponds to the puck position on the tablet.

ka-soru ga  taburetto de   no  pakku iti       ni  soutou     suru

cursor  NOM tablet     ATTR ADN puck  position DAT correspond pres


parsing: 36sec  parses: 4  deep: 1  transfer: 49sec  xltns: 2
translation 1


ka-soru ga  taburetto no  ue          no  pakku iti       ni  soutou     suru

cursor  NOM tablet    ADN above_place ADN puck  position DAT correspond pres

translation 2
```

Acknowledgements

The work on which this paper is based was supported by International Computers Limited (ICL) and by the UK Science and Engineering Research Council under the Alvey programme for research in Intelligent Knowledge-Based Systems. We are grateful to Jiri Jelinek for information on Aidtrans, to Peter Whitelock for his essential part in Ntran, and to Natsuko Holden and Elaine Pollard for comments on earlier drafts.

References

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

Gazdar, G., E. Klein, G. Pullum & I. Sag. 1985. Generalized Phrase Structure Grammar. Basil Blackwell, Oxford.

Jelinek, J. 1978. Integrated Japanese-English Grammar Dictionary. Sheffield.

Johnson, R. L. 1987. Translation. In Whitelock, P. J., M. McGee Wood, H. L. Somers, R. L. Johnson, & P. A. Bennett, eds. Linguistic Theory and Computer Applications. Academic Press, London.

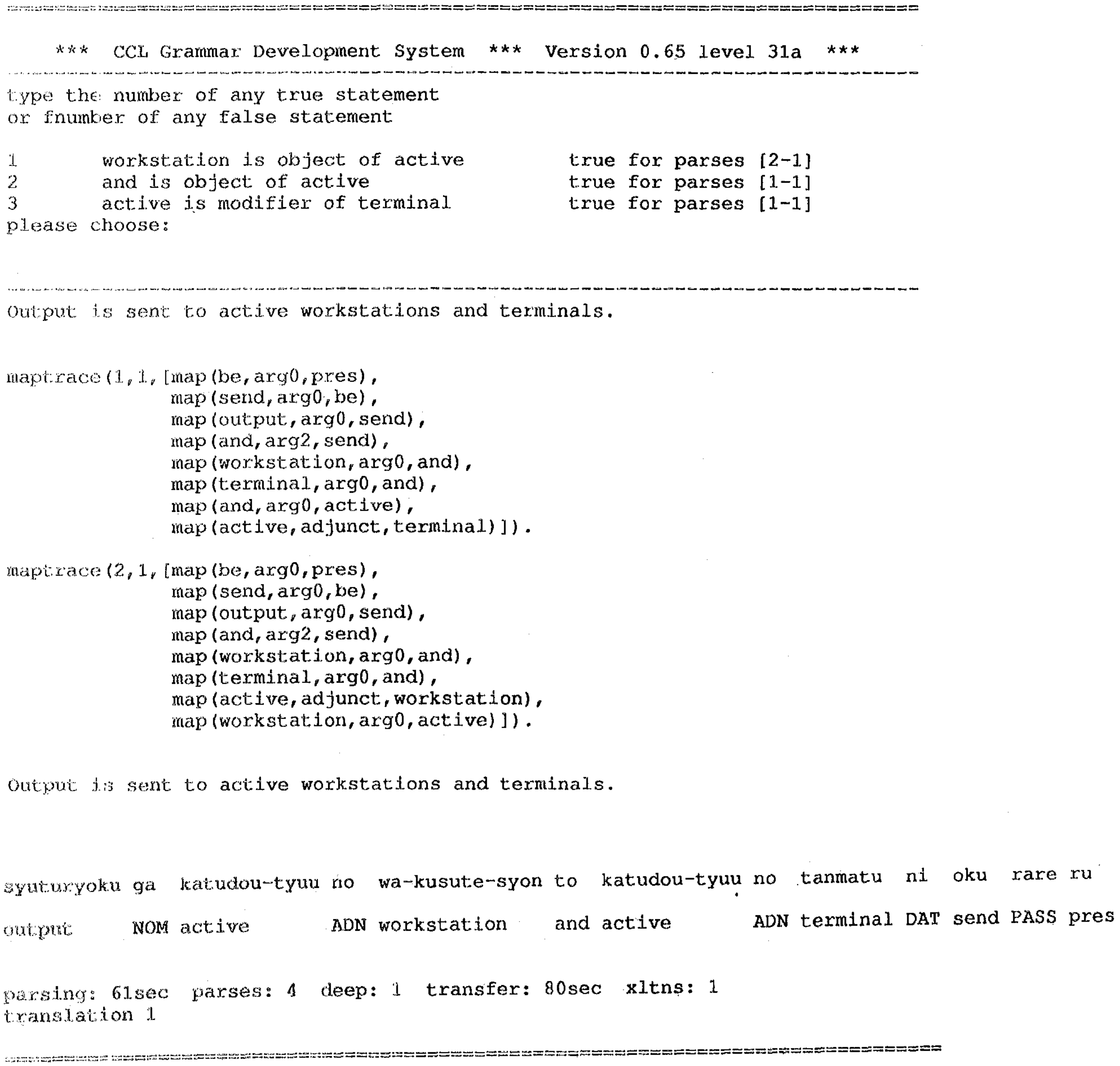
Johnson, R. L., & P. J. Whitelock. 1987. "Machine translation as an expert task". In Nirenburg, S., ed. Machine Translation. Cambridge University Press, Cambridge.

Steedman, M. J. 1985. "Dependency and Coordination in the Grammar of Dutch and English". Language.

Whitelock, P. J. 1988. "A Feature-based Categorial Morpho-syntax for Japanese". In Reyle, U. and C. Rohrer, eds. Natural Language Parsing and Linguistic Theories. Reidel, Dordrecht.

Whitelock, P. J., M. McGee Wood, B. J. Chandler, N. Holden, & H. J. Horsfall. 1986. "Strategies for Interactive Machine Translation". Proceedings of Coling86.

Wood, M. McGee, E. Pollard, H. J. Horsfall, N. Holden, B. J. Chandler, & J. J. Carroll. 1987. "Dictionary Organization for Machine Translation". Proceedings of ACL Europe 87.



```
     ***  CCL Grammar Development System  ***  Version 0.65 level 31a  ***

type the number of any true statement
or fnumber of any false statement

1       workstation is object of active         true for parses [2-1]
2       and is object of active                 true for parses [1-1]
3       active is modifier of terminal          true for parses [1-1]
please choose:


Output is sent to active workstations and terminals.


maptrace(1,1,[map(be,arg0,pres),
              map(send,arg0,be),
              map(output,arg0,send),
              map(and,arg2,send),
              map(workstation,arg0,and),
              map(terminal,arg0,and),
              map(and,arg0,active),
              map(active,adjunct,terminal)]).

maptrace(2,1,[map(be,arg0,pres),
              map(send,arg0,be),
              map(output,arg0,send),
              map(and,arg2,send),
              map(workstation,arg0,and),
              map(terminal,arg0,and),
              map(active,adjunct,workstation),
              map(workstation,arg0,active)]).


Output is sent to active workstations and terminals.



syuturyoku ga  katudou-tyuu no  wa-kusute-syon to  katudou-tyuu no  tanmatu  ni  oku  rare ru

output     NOM active       ADN workstation    and active        ADN terminal DAT send PASS pres


parsing: 61sec  parses: 4  deep: 1  transfer: 80sec  xltns: 1
translation 1
```