GETTING IDIOMS INTO A LEXICON BASED PARSER'S HEAD

Oliviero Stock
I.P. - Consiglio Nazionale delle Ricerche
Via dei Monti Tiburtini 509
00157 Roma, Italy

ABSTRACT

An account is given of flexible idiom processing within a lexicon based parser. The view is a compositional one. The parser's behaviour is basically the "literal" one, unless a certain threshold is crossed by the weight of a particular idiom. A new process will then be added. The parser, besides yielding all idiomatic and literal interpretations embodies some claims of human processing simulation.

1. Motivation and comparison with other approaches

Idioms are a pervasive phenomenon in natural languages. For instance, the first page of this paper (even if written by a non-native speaker) includes no less than half dozen of them. Linguists have proposed different accounts for idioms, which are derived from two basic points of view: one point of view considers idioms as the basic units of language, with holistic characteristics, perhaps including words as a particular case; the other point of view emphasizes instead the fact that idioms are made up of normal parts of speech, that play a precise role in the complete idiom. An explicit statement within this approach is the Principle of Decompositionality (Wasow, Sag and Nunberg 1982): "When an expression admits analysis as morphologically or syntactically complex, assume as an operating hypothesis that the sense of the expression arises from the composition of the senses of its constituent parts". The syntactic consequence is that idioms are not a different thing from "normal" forms.

Our view is of the latter kind. We are aware of the fact that the flexibility of an idiom, depends on how recognizable its metaphorical origin is. Within flexible word order languages the flexibility of idioms seems to be even more closely linked to the strengths of particular syntactic constructions.

Let us now briefly discuss some computational approaches to idiom understanding. Applied computational systems must necessarily have a capacity for analyzing idioms. In some systems there is a preprocessor delegated to the recognition of idiomatic forms. This preprocessor replaces the group of words that make for one idiom with the word or words that convey the meaning involved. In ATN systems instead, specially if oriented towards a particular domain, sometimes there are sequences of particular arcs inserted in the network, which, if transited, lead to the recognition of a particular idiom (e.g. PLANES, Waltz 1978). LIFER (Hendrix 1977), one of the most successful applied systems, was based on a semantic grammar, and within this mechanism idiom recognition was easy to implement, without considering flexibility. Of course, in all these systems there is no intention to give an account of human processing. PHRAN (Wilensky and Arens 1980) is a system based entirely on pattern recognition. Idiom recognition, following Fillmore's view (Fillmore 1979) is considered the basic resource all the way down to replace the concept of grammar based parsing. PHRAN is based on a data base of patterns (including single words, at the same level), and proceeds deterministically, applying the two principles "when in doubt choose the more specific pattern" and "choose the longest pattern". The limits of this approach lie in the capacity of generating various alternative interpretations in case of ambiguity and in running the risk of having an eccessive spread of nonterminal symbols if the data base of idioms is large. A recent work on idioms with a similar perspective is Dyer and Zernik (1986).

The approach we have followed is different. The goals we had with our work must be stated explicitly: 1) to yield a cognitive model of idiom processing; 2) to integrate

idioms in our lexical data, just as further information concerning words (as in a traditional dictionary) 3) to insert all this in the framework of WEDNESDAY 2 (Stock 1986), a nondeterministic lexicon based parser.

To anticipate the cognitive solution we are discussing here: idiom understanding is based on normal syntactic analysis with word driven recognition in the background. When a certain threshold is crossed by the weight of a particular idiom, the latter starts a process of its own, that may eventually lead to a complete interpretation.

Some of the questions we have dealt with are: how are idioms to be specified? b) when are they recognized? c) what happens when they are recognized? d) what happens afterwards?

2. A summary of WEDNESDAY 2

WEDNESDAY 2 (Stock 1986) is a parser based on linguistic knowledge distributed fundamentally through the lexicon. The general viewpoint of the linguistic representation is not far from LFG (Kaplan & Bresnan 1982), although independently conceived.

A word interpretation includes:

- a semantic representation of the word, in the form of a semantic net shred;

- static syntactic information, including the category, features, indication of linguistic functions that are bound to particular nodes in the net. One particular specification is the Main node, the head of the syntactic constituent the word occurs in;

- dynamic syntactic information, including impulses to connect pieces of semantic information, guided by syntactic constraints. Impulses look for "fillers" on a given search space. They have alternatives, (for instance the word *tell* has an impulse to merge its object node with the Main node of either an NP or a subordinate clause). An alternative includes: a contextual condition of applicability, a category, features, marking, side effects (through which, for example, coreference between subject of a subordinate clause and a function of the main clause can be indicated). Impulses may also be directed to a different search space than the normal one with a mechanism that can deal with long distance dependencies;

- measures of likelihood. These are measures that are used in order to derive an overall measure of likelihood of a partial analysis. Measures are included for the likelihood of that particular reading of the word and for aspects attached to an impulse: a) for one particular alternative b) for the relative position the filler c) for the overall necessity of finding a filler.

- a characterization of idioms involving that word (see next paragraph).

The only other data that the parser uses are in the form of simple (non augmented) transition networks that only provide restrictions on search spaces where impulses can look for fillers. In more traditional words these networks deal with the distribution of constituents. A distinguished symbol, $EXP, indicates that only the occurrence of something expected by preceding words (i.e. for which an impulse was set up) will allow the transition. It is stressed that inside a constituent the position of elements can be free. In WEDNESDAY 2 one can specify in a natural and nonredundant way, all the graduality from obligatory positions, to obligatory precedences to simple likelihoods of relative positions.

The parser is based on an extension of the idea of chart parsing [Kay 1980, Kaplan 1973] [see Stock 1986]. What is relevant here is the fact that "edges" correspond to search spaces. They are complex data structures provided with a rich amount of information including a semantic interpretation of the fragment, syntactic data, pending impulses, an overall measure of likelihood etc. Data on an edge are "unified" dynamically.

Parsing goes basically bottom-up with top-down confirmation, improving the so called Left Corner technique. When a lexical edge with category C is added to the chart, its First Left Cross References F(C) are fetched. First Left Cross References are defined recursively: for every lexical category C, the set of initial states that allow for transitions on C, or the set of initial states (without repetitions) that allow for transitions on symbols in F(C). So, for instance, F(Det) = {NP,S}, at least.

For each element in F(C) an edge of a special kind is added to the chart. These special edges are called *sleeping edges*. A sleeping edge at a vertex V_s is *awakened*, i.e. causes the introduction of a normal active edge iff there is an active edge arriving at Vs that may be extended with an edge with the category of S. If they are not awakened, sleeping edges play no role at all in the process.

An agenda is provided which includes tasks of several different types, including *lexical tasks*, *extension tasks*, *insertion tasks* and *virtual tasks*. A lexical task specifies

a possible reading of a word to be introduced in the chart as an inactive edge. An extension task specifies an active edge and an inactive edge that can extend it (together with some more information). An insertion task specifies a nondeterministic unification operation. A virtual task consists in extending an active edge with an edge displaced to another point of the sentence, according to the mechanism that treats long distance dependencies. At each stage the next task chosen for execution is the value of a scheduling-selecting function.

The parser works asymmetrically with respects to the "arrival" of the Main node: before the Main node arrives, an extension of an edge causes almost nothing. On the arrival of the Main, all the candidate fillers must find a compatible impulse and all impulses concerning the main node must find satisfaction. If all this does not happen then the new edge supposedly to be added to the chart is not added: the situation is recognized as a failure. After the arrival of the Main, each new head must find an impulse to merge with , and each incoming impulse must find satisfaction. Again, if all this does not happen, the new edge will not be added to the chart.

Dynamically, apart from the general behaviour of the parser, there are some particular restrictions for its nondeterministic behaviour, that put into effect syntax-based dynamic disambiguation.

1) the $EXP arc allows for a transition only if the configuration in the active edge includes an impulse to link with the Main of the proposed inactive edge.

2) The sleeping edge mechanism prevents edges not compatible with the left context from being established.

3) A search space can be closed only if no impulse that was specified as having to be satisfied remains. In other words, if in a state with an outgoing EXIT arc, an active edge can cause the establishing of an inactive edge only if there are no obligatory impulses left.

4) A proposed new edge A' with a verb tense not matching the expected values causes a failure, i.e. that A' will not be introduced in the chart.

5) Failure is caused by inadequate mergings, with relation to the presence, absence or ongoing introduction of the Main node.

Comparing to the criteria established for LFG for functional compatibility of an f-structure [Kaplan & Bresnan 1982], the following can be said of the dynamics outlined here. *Incompleteness* recognition performs as specified in 3), and furthermore there is an earlier check when the Main arrives, in case there were obligatory impulses to be satisfied at that point (e.g. an argument that must occur before the Main). *Incoherence* is completely avoided after the Main has arrived, by the $EXP arc mechanism; before this point, it is recognized as specified in 5) above, and causes an immediate failure. *Inconsistency* is detected as indicated in 4) and 5). As far as 5) is concerned, though, the attitude is to "activate" impulses when the right premises are present and to "look for the right thing" and not to "check if what was done is consistent".

Note that a morphological analyzer, WED-MORPH, linked to WEDNESDAY 2, plays a substantial role, specially if the language is Italian. In Italian you may find words like *rifacendogliene*, that stands for *while making some (of them) for him again*. The morphological analyzer not only recognizes complex forms, but must be able to put together complex constraints originated in part by the stem and in part by the affixes. The same holds for the semantic representation and will have consequences in our dealing with idioms. Fig. 1 shows a diagram of WEDNESDAY 2

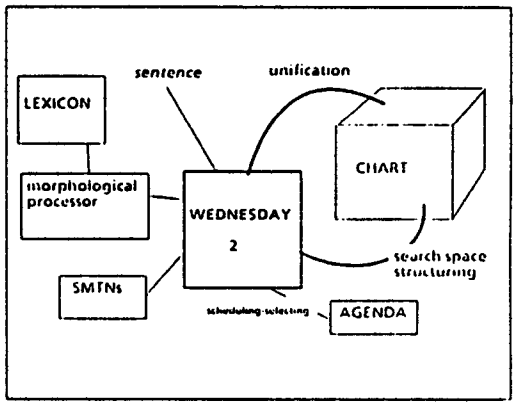


Fig. 1

3. Specification of idioms in the lexicon

Idioms are introduced in the lexicon as further specifications of words, just as in a normal dictionary. They may be of two types: a) canned phrases, that just behave as several-word entries in the lexicon (there is nothing particularly interesting in that, so we shall not go into detail here); b) flexible idioms; these idioms are

described in the lexicon bound to the particular word representing the "thread" of that idiom; in WEDNESDAY 2 terms, this is the word that bears the Main of the immediate constituent including the idiom. Thus, if we have an idiom like *to build castles in the air*, it will be described along with the verb, *to build*.

After the normal word specifications, the word may include a list of idiomatic entries. Fig.2 shows a BNF specification of idioms in the lexicon. The symbol + stands for "at least one occurrence of what precedes"). Each idiom is described in two sections: the first one describes the elements that characterize that idiom, expressed coherently with the normal characterization of the word, the second one describes the interpretation, i.e. which substitutions should be performed when the idiom is recognized.

Let us briefly describe Fig. 2. The lexicalform indicates whether passivization (that in our theory, like in LFG, is treated in the lexicon) is admitted in the idiomatic reading. The idiom-stats, describing configurations of the components of an idiom, are based on the basic impulses included in the word. In other words constituents of an idiom are described as particular fillers of linguistic functions or particular modifiers. For example *build castles in the air*, when *build* is in an active form, has *castles* as a further description of the filler of the OBJ function and the string *in the air* as a further specification of a particular modifier that may be attached to the Main node. MORESPECIFIC, the further specification of an impulse to set a filler for a function includes: a reference to one of the possible alternative types of fillers specified in the normal impulse, a specification that describes the fragment that is to play this particular role in the idiom, and the weight that this component has in the overall recognition of the idiom. IDMODIFIER is a specification of a modifier, including the description of the fragment and the weight of this component. CHANGEIMPULSE and REMOVEIMPULSE consent an alteration of the normal syntactic behaviour. The former specifies a new alternative for a filler for an existing function, including the description of the component and its weight (for instance the new alternative may be a partial NP instead of a complete NP (as in *take care*), or a NP marked differently from usual). The latter specifies that a certain impulse, specified for the word, is to be considered to have been removed for this idiom description.

There are a number of possible fragment specifications, including string patterns, semantic patterns, morphological variations, coreferences etc.

Substitutions include the semantics of the idiom, which are supposed to take the place of the literal semantics, plus the specfication of the new Main and of the bindings for the functions. New bindings may be included to specify new semantic linkings not present in the literal meaning (e.g. *take care of <someone>*, if the meaning is *to attend to <someone>*, then *<someone>* must become an argument of *attend*).

```
<idioms>::=(IDIOMS <idiomentry>+)
<idiomentry>::=(<lexicalform><idiom-stat>+ SUBSTITUTIONS <idiomsubst>+)
<lexicalform>::=T/(NOT-PASSIVE)
<idiom-stat>::=(MORESPECIFIC<lingfunc><alternnum><fragmentspec><weight>)/
                (CHANGEIMPULSE<lingfunc><alternative>+<fragmentspec><weight>)/
                (IDMODIFIER <fragmentspec><weight>)/
                (REMOVEIMPULSE <lingfunc>)
<alternative>::=(<test><fillertype><beforelh><features><mark><sideffect><fragmentspec>)
<fragmentspec>::=(WORD<word>)/(FIXWORDS<wordseq>)/(FIRSTWORDS <wordseq>)/
                  (MORPHWORD <wordroot>)/(SEM (<concept>+)<prep>)/(EQSUBJ)
<idiomsubst>::=(SEM-UNITS<sem-unit>+)/(MAIN<node>)/
                (BINDINGS(<lingfunc><node>)+)/
                (NEWBINDINGS(<node><lingfunc path>)+)
```

Fig. 2

4. Idiom processing

Idiom processing works in WEDNESDAY 2 integrated in the nondeterministic, multiprocessing-based behaviour of the parser. As the normal (literal) analysis proceeds and partial representations are built, impulses are monitored in the background, checking for possible idiomatic fragments. Monitoring is carried on only for fragments of idioms not in contrast with the present configuration. A dynamic activation table is introduced with the occurrence of a word that has some idiom specification associated. Occurrence of an expected fragment of an idiom in the table raises the level of activation of that idiom, in proportion to the relative weight of the fragment. If the configuration of the sentence contrasts with one fragment then the relative idiom is discarded from the table. So all the normal processing goes on, including the possible nondeterministic choices, the establishing of new processes etc. The activation tables are included in the edges of the chart.

When the activation level of a particular idiom crosses a fixed threshold, a new process is introduced, dedicated to that particular idiom. In that process, only that, idiomatic interpretation is considered. Thus, in the first place, an edge is introduced, in which substitutions are carried on; the process will proceed with the idiomatic representation. Note that the process begins at that precise point, with all the previous literal analysis acquired to the idiomatic analysis. The original process goes on as well (unless the fragment that caused the new process is non syntactic and only peculiar to that idiom); only, the idiom is removed from the active idiom table. At this point there are two working processes and it is a matter of the (external) scheduling function to decide priorities. What is relevant is: a) still, the idiomatic process may result in a failure: further analysis may not confirm what has been hypothesized as an idiom; b) a different idiomatic process may be parted from the literal process at a later stage, when its own activation level crosses the threshold.

Altogether, this yields all the analyses, literal and idiomatic, with likelihoods for the different interpretations In addition, it seems a reasonable model of how humans process idioms. Some psycholinguistic experiments have supported this view (Cacciari & Stock, in preparation) which is also compatible with the model presented by Swinney and Cutler (1978).

Here we have disregarded the situation in which a possible idiomatic form occurs and its role in disambiguating. The whole parsing mechanism in WEDNESDAY 2 is based on dynamic unification, i.e. at every step in the parsing process a partial interpretation is provided; dynamic choices are performed scheduling the agenda on the base of the relation between partial interpretations and the context.

5. An example

As an example let us consider the Italian idiom *prendere il toro per le corna* (literally: *to take the bull by the horns*; idiomatically: *to confront a difficult situation*). The verb *prendere* (*to take*) in the lexicon includes some descriptions of idioms. Fig. 3 shows the representation of *prendere* in the lexicon. The stem representation will be unified with other information and constraints coming from the affixes involved in a particular form of the verb. The first portion of the representation is devoted to the literal interpretation of the word, and includes the semantic representation, the likelihood of that reading, and functional information, included the specification of impulses for unification. The numbers are likelihoods of the presence of an argument or of a relative position of an argument. The

```
(sem-units (n1(p-take n2 n3)))
(likeliradix 0.8)
(main n1)
(lingfunctions (subj n2)(obj n3))
(cat v)
(uni (subj)
     (must 0.7)
     ((t np  0.9 nil nom)))
(uni (obj)
     (must)
     ((t np 0.3 nil acc)))
(idioms ((t
          (morespecific (obj) 1 (fixwords il toro) 8)
          (idmodifier (fixwords per le corna) 10)
          substitutions
          (sem-units (m1(p-confront m2 m3))
                     (m4 (p-situation m3))
                     (m5 (p-difficult m3)))
          (main m1)
          (bindings (subj m2))]
```

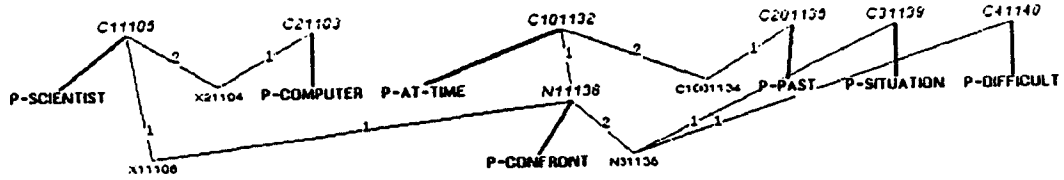
Fig. 3

second portion, after "idioms" includes the idioms involving "prendere". In Fig. 3 only one such idiom is specified. It is indicated that the idiom can also occur in a passive form and the specification of the expected fragments is given. The numbers here are the weights of the fragments (the threshold is fixed to 10). The substitutions include the new semantic representation, with the specification of the main node and of the binding of the subject. Note that the surface functional representation will not be destroyed after the substitutions, only the semantic (logical) representation will be recomputed, imposing its own bindings.

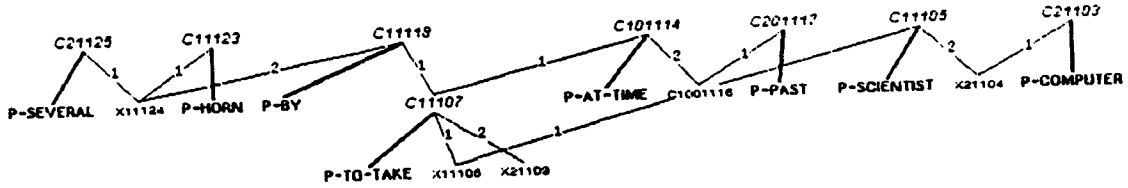
As mentioned, Italian allows great flexibility. Let the input sentence be *l'informatico prese per le corna la capra* (literally: *the computer scientist took by the horns the goat*). When *prese* (*took*) is analyzed its idiom activation table is inserted. When the modifier *per le corna* (*by the horns*) shows up, the activation of the idiom referred to above crosses the threshold (the sum of the two weights goes up to 12). A new process starts at this point, with the new interpretation unified with the previous interpretation of the Subject. Also, semantic specifications coming from the suffixes are reused in the new partial interpretation. The process just departs from the literal process, no backtracking is performed. At this point we have two processes going on: an idiomatic process, where the interpretation is already *the computer scientist is confronting a difficult situation* and a literal process, where, in the background, still other active idioms monitor the events. In fig. 4 the two semantic representations, in the form of semantic networks, are shown. When the last NP, *la capra* (*the goat*), is recognized, the idiomatic process fails (it needed *the bull* as Object). The literal process yields its analysis, but, also, another idiom crosses the threshold, starts its process with the substitutions and immediately concludes positively. This latter, unlikely, idiomatic interpretation means *the computer scientist confused the goat and the horns*.

6. Implementation

WEDNESDAY 2 is implemented in Interlisp-D and runs on a Xerox 1186. The idiom recognition ability was easily integrated into the system. The performance is very satisfying, in particular with regard to the flexibility present in Italian. Around the parser a rich environment has been built. Besides allowing easy editing and graphic inspecting of resulting structures, it allows interaction with the agenda and exploration of heuristics in order to drive the multiprocessing mechanism of WEDNESDAY 2.



a)



b)

Fig. 4

This environment constitutes a basic resource for exploring cognitive aspects, complementary to laboratory experiments with humans.

At present we are also working on an implementation of a generator that includes the ability to produce idioms, based on the same data structure and principles as the parser.

Acknowledgements

Thanks to Cristina Cacciari for many discussions and to Federico Cecconi for his continuous help.

References

Dyer, M. & Zernik, U. Encoding and Acquiring Meaning for Figurative Phrases. In *Proceedings of the 24th Meeting of the Association for Computational Linguistics*. New York (1986)

Fillmore, C. Innocence: a Second Idealization for Linguistics. In *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society*. University of California at Berkeley, 63-76 (1979).

Hendrix, G.G. LIFER: a Natural Language Interface Facility. *SIGART Newsletter* Vol. 61 (1977).

Kaplan, R. A general syntactic processor. In Rustin, R. (Ed.), *Natural Language Processing*. Englewood Cliffs, N.J.: Prentice-Hall (1973)

Kaplan,R. & Bresnan,J. Lexical-Functional Grammar: a formal system for grammatical representation. In Bresnan,J., Ed. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, 173-281 (1982)

Kay, M. Algorithm Schemata and Data Structures in Syntactic Processing. Report CSL-80-12, Xerox, Palo Alto Research Center, Palo Alto (1980)

Stock, O. Dynamic Unification in Lexically Based Parsing. In *Proceedings of the Seventh European Conference on Artificial Intelligence*. Brighton, 212-221 (1986)

Swinney, D.A., & Cutler, A. The Access and Processing of Idiomatic Expressions. *Journal of Verbal Learning and Verbal Behaviour*, 18, 523-534 (1978)

Waltz, D. An English Language Question Answering System for a Large Relational Database. *Communications of the of the Association for Computing Machinery*, Vol. 21, N. 7 (1978).

Wasow, T., Sag, I., Nunberg, G. Idioms: an interim report. *Preprints of the International Congress of Linguistics*, 87-96, Tokyo (1982)

Wilensky, R. & Arens, Y. PHRAN - A Knowledge Based Approach to Natural Language Analysis. University of California at Berkeley, ERL Memorandum No. UCB/ERL M80/34 (1980).