Ambiguity resolution and the retrieval of idioms: two approaches.

Erik-Jan van der Linden Wessel Kraaij

Institute for Language Technology and AI
Tilburg University
PO box 90153
5000 LE Tilburg
The Netherlands
E-mail: vdlinden@kub.nl

Abstract

When an idiomatic expression is encountered during natural language processing, the ambiguity between its idiomatic and non-idiomatic meaning has to be resolved. Rather than including both meanings in further processing, a *conventionality-principle* could be applied. This results in best-first processing of the idiomatic analysis. Two models are discussed fot the lexical representation of idioms. One extends the notion *continuation class* from two-level morphology, the other is a localist, connectionist model. The connectionist model has an important advantage over the continuation class model: the conventionality principle follows naturally from the architecture of the connectionist model.

Keywords: idiom processing, ambiguity resulution, two-level morphology, connectionism.

1 Introduction

In this paper we discuss the resolution of the ambiguity between the non-idiomatic and the idiomatic reading of a phrase that is possibly idiomatic. A choice between these readings can be made using various kinds of linguistic information, but we claim that it can be made on the basis of the mere fact that one of the analyses is idiomatic, and that this choice does not have to be stipulated explicitly, but follows naturally from the architecture of the lexicon and the retrieval process, if an appropriate model of the lexicon is used.

In section (2), we firstly state our approach to natural language processing (NLP) in general. Next (3), we document the claim that idioms should be stored in the lexicon as holistic lexical units, and discuss the processing of idioms (4). Then we present two models for the lexical representation and retrieval of idioms: one extends the notion *continuation class* of the two-level model (Koskenniemi 1983) (5); the other is a simple localist connectionist model (6).

In this paper we limit ourselves to aspects relevant for our 'lexical' approach. We will not discuss such issues as *semantic decomposability* (Gibbs and Nayak 1989) of idioms (but see van der Linden 1989, for a model of incremental syntactic/semantic processing of idioms in Categorial Grammar, and a more elaborate discussion of the issues mentioned in (3.1)).

2 General approach to NLP

In case of ambiguity, the NL processor has to choose between a number of possible analyses. In order to determine which one is most likely, *all* analyses can be examined using a *breadth-first* or *depth-first* strategy. However, this results in a time and space consuming process. Incremental *best-first* examination of 'promising' analyses seems more appropriate, and seems to be a property of human language processing (cf. Thibadeau, Just and Carpenter 1982). However, this approach leaves us with the question what criteria should be applied to select one of the analyses for further examination. In this paper we examine one aspect of this general question, namely the choice between an idiomatic and a non-idiomatic analysis.

3 Idioms and the lexicon

3.1 Linguistics

Within the literature 'Traditional wisdom dictates that an idiom is by definition a constituent or series of constituents where interpretation is not a compositional function of the interpretation of its parts.' (Gazdar et al. 1985, p 327). [1] Rather than giving a definition of idioms that states what the meaning of an idiom *isn't*, we prefer a definition that states what the meaning *is* (van der Linden 1989).

> Idioms are multi-lexemic expressions, the meaning of which is a property of the whole expression.

[1] Cf. Wood (1986): an idiom is 'wholly non-compositional in meaning'.

Some attempts have, however, been made to assign the meaning of an idiom to the parts of the idiom. These can roughly be divided in two:
- assignment of the idiomatic meaning of the expression to one of the parts of the idiom, and *no* meaning to the other part (Ruhl, cited in Wood 1986). In the case of *kick the bucket*, the meaning **die** is assigned to *kick*, and *no* meaning to the other part. This raises the question however, why one cannot say *Pat rested the bucket* to mean *Pat rested* (Wasow et al. 1983).
- assignment of idiomatic interpretations to all parts. Compositional combination of these meanings results in an idiomatic meaning for the whole expression (Gazdar et al. 1985). This analysis has a number of problems. Gazdar et al. use *partial functions* to avoid combination of an idiomatic functor with a non-idiomatic argument, but do not explain how to avoid combination of a non-idiomatic functor with an idiomatic argument. In our view this can only be solved by introducing partial arguments, to our knowledge a non-existing notion, or by accepting that *all* functors are partial, which is not common in linguistics.
We conclude that the meaning of an idiom is a property of the whole expression, and should be represented in the lexicon. [2]

3.2 Psycholinguistics

The same observation arises from psycholinguistic research. Idioms are stored and accessed as lexical items, not from some special list that is distinct from the lexicon[3] (Swinney and Cutler 1979). Furthermore, idioms are stored as holistic entries in the mental lexicon (Swinney and Cutler 1979; Lancker and Canter 1980; Lancker, Canter and Terbeek 1981; cf. the notion 'configuration' in Cacciari and Tabossi 1988).

4 Processing idioms

Phrases consisting of idioms *can* in most cases be interpreted non-idiomatically as well. [4] It has however frequently been observed that very rarely an idiomatic phrase *should* in fact be interpreted non-idiomatically (Koller 1977, p. 13; Chafe 1968, p. 123; Gross 1984, p. 278; Swinney 1981, p. 208). Also, psycholinguistic research indicates that in case of an ambiguity, there is clear preference for the idiomatic reading (Gibbs 1980; Schweigert and Moates 1988). We will refer to the fact that phrases should be interpreted according to the idiomatic, non-compositional, lexical, conventional, meaning, as the 'conventionality' principle. [5] If this principle could be modeled in an appropriate way, this would be of considerable help in dealing with idioms: as soon as the idiom has been identified, the ambiguity can be resolved and 'higher' processes do not have to examine the various analyses.
There is one more issue, that requires some consideration: when can and does an incremental processor *start* looking for idioms? From psycholinguistic research it appears that idioms are not activated when the 'first' (content) word is encountered (Swinney and Cutler 1979). There is, from the computational point of view, no need to start 'looking' for idioms, when only the first word has been found [6]: that would result in increase of the processing load at higher levels.

Stock In Stock's (1989) approach to ambiguity resolution, firstly, the idiomatic and the non-idiomatic analysis are processed in parallel. An external scheduling function gives priority to one of these analyses. Secondly, Stock starts looking for idioms when the 'first' word has been encountered. As we have stated, both increase the load on higher processes.

5 An extension of the notion *continuation class*

Lexical representation Lexical entries in two-level morphology are represented in a trie structure, which enables incremental lookup of strings. A lexical entry consists of a lexical representation, linguistic information, and a so-called continuation class, which is a list of sublexicons "the members of which may follow" (Koskenniemi 1983, p. 29) the lexical entry. In the continuation class of an adjective, one could for instance find a reference to a sublexicon containing comparative endings (ibid. p. 57). An obvious extension is to apply this notion beyond the boundaries of the word. A continuation class of an entry A could contain references to the entries that form an idiom with A. An example is (1a) (Note that we use a graphemic code for the lexical representations).

(1)

(a)

```
k-i-c-k*---b-u-c-k-e-t*
          h-a-b-i-t*
           \e-e-l-s*
```

[2] Although other opinions exist (Pesetsky 1985).

[3] As has been defended by Bobrow and Bell (1973).

[4] Exceptions are idioms that contain words that occur in idioms only (*spin and span, queer the pitch*), and ungrammatical idioms (*trip the light fantastic*).

[5] The same can be observed for compounds: these are not interpreted compositionally, but according to the lexical, conventional meaning (Swinney 1981).

[6] One might argue that words that only occur in idioms could activate the idioms when encountered as 'first' word. There is however only psycholinguistic evidence that this occurs in a highly semantically determined contexts (Cacciari and Tabossi 1988).

(b)

```
DO read a letter
   IF word has been found THEN
     IF this word forms an idiom
        with previous word(s)
     THEN make idiom information
          available to syn/sem process
     ELSE make word information
          available to syn/sem process
UNTIL no more letters in input.
```

Algorithm A simple algorithm is used to find idioms (in (1b) the relevant fragment of the algorithm is represented in pseudocode). The result of the application of the algorithm is that linguistic information associated with the idioms is supplied to the syntactic/semantic processor. The linguistic information includes the precise form of the idiom, the possibilities for modification etc. (cf. van der Linden 1989).
A toy implementation of the lexicon structure and the algorithm has been made in C.

6 A connectionist model

The second model we present here for the lexical representation and retrieval of idioms is an extension of a simple localist connectionist model for the resolution of lexical ambiguity (Cottrell 1988 [7]). The model (2) consists of four levels. Units at the lowest level represent the smallest units of form. These units activate units on the level that represents syntactic discriminations, which in turn activate units on the semantic level. The semantic features activate relational nodes in the semantic network. *Within* levels, inhibitory links may occur; *between* levels excitatory links may exist. However, there are no inhibitory links within the semantic network.
The meaning of *idioms* is represented as all other relational nodes in the semantic network. On the level of semantic features, the idiom is represented by a unit that has a *gate* function similar to so-called *SIGMA-PI units* (Rumelhart and McClelland 1986, p. 73): in order for such a unit A to receive activation, all units excitating A bottom-up should be active. If one of the units connected to a unit A is not active, A does not receive activation. Thus when the first word of an idiom is encountered, the idiom is not activated, because the other word(s) is (are) *not* active. However, once *all* relevant lexemes have been encountered in the input, it becomes active. Note that an external syntactic module excitates one of the nodes in case of syntactic ambiguity. Since there is more than one syntactic unit activating the idiom, the overall activation of the idiom becomes higher than competing nodes representing non-idiomatic meanings. Or, to put it differently, the idiom represents the simplest hypothesis that accounts for the meaning of the lexemes in the input. The idiom is the strongest competitor, and inhibits the non-idiomatic readings.
Without need for feedback from *outside* the model, the conventionality principle is thus modeled as a natural consequence of the architecture of the model. The connectionist model has been implemented in C with the use of the Rochester Connectionist Simulator (Goddard et al. 1989). In the appendix we give a description of technical details of this implementation.

[7] For an introduction to connectionist models, see Rumelhart and McClelland (1986); for a critical evaluation see Fodor and Phylyshyn (1988).

7 Concluding remarks

Ambiguity in the case of idiomatic phrases can be resolved on the basis of the conventionality principle. When compared to strategies as used in Stock (1989), both models presented here have the advantage that they don't process idioms until all relevant lexical material has been encountered in the input and operate in a *best-first* fashion. Therefore they contribute to the efficiency of the parsing process. The connectionist model has one further advantage: the conventionality principle results naturally from the architecture of the model, and does not have to be stipulated explicitely. The obvious disadvantage of this model is the necessity for parallel hardware for realistic implementations. Future research will include the true integration with a syntactic module. Then, it will also be able to take the precise syntactic form of idiomatic phrases into account.

Acknowledgements

The authors would like to thank Walter Daelemans and the visitors of the Colloquium "Computer and Lexicon" 12 & 13 Oct, 1989, Utrecht for their useful comments. Wietske Sijtsma commented on English grammaticality and style.

References

Bobrow, S. and Bell, S. 1973. On catching on to idiomatic expressions. *Memory and Cognition* 3, 343-346.

Cacciari, C. and Tabossi, P. 1988. The comprehension of Idioms. *Journal of Memory and Language*, 27, 668-683.

Chafe, W. 1968. Idiomaticity as an anomaly in the Chomskyan Paradigm. *Foundations of Language* 4, 109-127.

Cottrell, G. 1989 A model of lexical access of ambiguous words. In: Small, S., Cottrel, G., and Tanenhaus, M., 1988. *Lexical ambiguity resolution.* San Mateo: Kaufmann.

Fodor, J. and Pylyshyn, Z. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 1988, 1-2, p. 3- 70

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

Gibbs, R., 1980. Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition* 8, 149-156.

Gibbs, R., and Nayak N. 1989. Psycholinguistic Studies on the syntactic behavior of Idioms. *Cognitive Psychology* 21, 100-138.

Goddard, N., Lynne, K., Mintz, T., and Bukys, L. 1989 Rochester Connectionist Simulator. Technical Report. University of Rochester.

Gross, M., 1984. Lexicon-grammar and the syntactic analysis of French. In *Proceedings COLING '84.*

Koller, W., 1977. *Redensarten: linguistische Aspecte, Vorkommensanalysen, Sprachspiel.* Tübingen: Niemeyer.

Koskenniemi, K. 1983. *Two-level morphology.* PhD-thesis. University of Helsinki.

Lancker, D. van and Canter, G. 1981. Idiomatic versus literal interpretations of ditropically ambiguous sentences. *Journal of Speech and Hearing Research* 24, 64-69.

Linden, E. van der, 1989. Idioms and flexible categorial grammar. In Everaert and van der Linden (Eds.) 1989.*Proceedings of the First Tilburg Workshop on Idioms.* Tilburg, the Netherlands, May 19, 1989. Tilburg: ITK.

Pesetsky, D. 1985. Morphology and Logical form. *Linguistic Inquiry*, 16 193-246.

Rumelhart, D. and McClelland, J. 1986. *Parallel Distributed processing. Explorations in the microstucture of cognition.* Cambridge, Massachusetts, MIT press.

Schweigert, W. and Moates, D. 1988. Familiar idiom comprehension. *Journal of Psycholinguistic Research*, 17, pp. 281-296.

Stock, O. 1989. Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind. *Computational Linguistics* 1, 1-19.

Swinney, D. 1981. Lexical processing during sentence comprehension: effects of higher order constraints and implications for representation. In: T. Meyers, J. Laver and J. Anderson (eds.) *The cognitive representation of speech.* North-Holland.

Swiney, D. and Cutler, A. 1979. The access and processing of idiomatic expressions. *JVLVB* 18, 523-534.

Thibadeau, R., Just, M., and Carpenter, P. 1982. A Model of the Time Course and Content of Reading. *Cognitive Science* 6, 157-203.

Wasow, T., Sag, I., and Nunberg, G. 1983. Idioms: an interim report. In Shiro Hattori and Kazuko Inoue (Eds.) *Proceedings of the XIIIth international congress of linguists.* Tokyo: CIPL 102-115.

Wood, M. McGee, 1986. A definition of idiom. Masters thesis, University of Manchester (1981). Reproduced by the Indiana University Linguistics Club.

Fig (2) Network representation

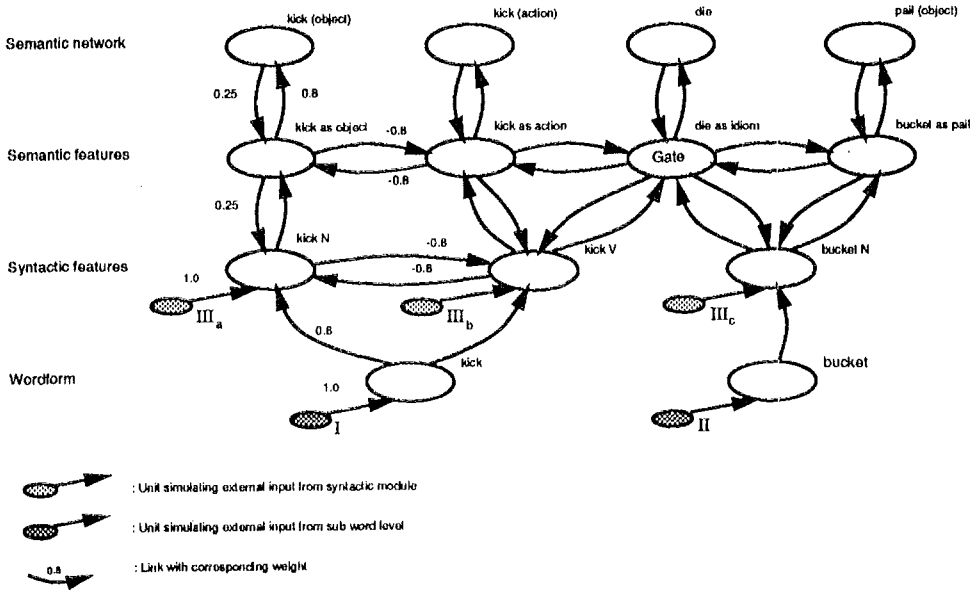


Fig (3) Unit structure

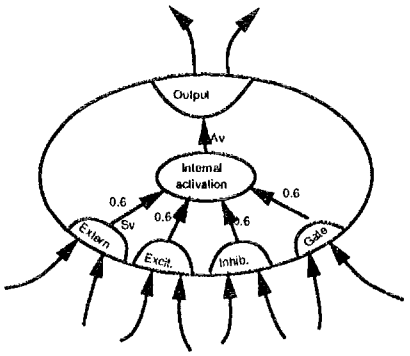


Fig (4) Activation level of the wordform and syntactic units

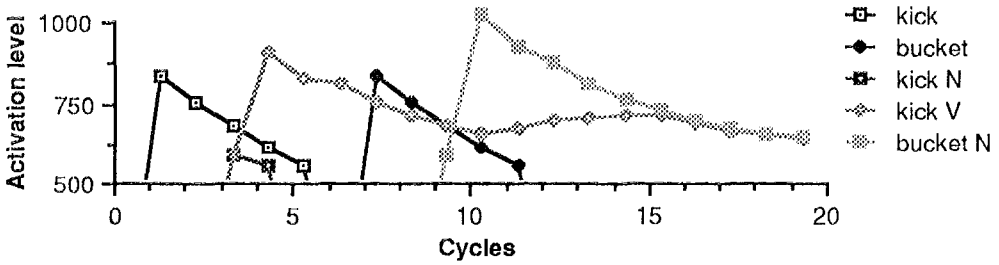
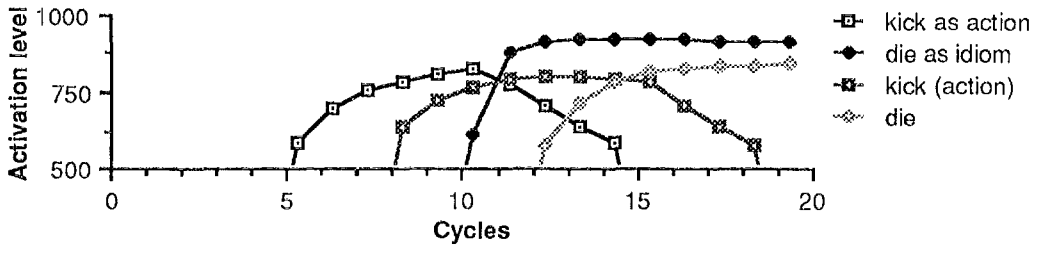


Fig (5) Activation level of the semantic units and the semantic network



Appendix

The connectionist model for the retrieval of idioms as presented in section 6 is based on the mechanism of interactive activation and competition (IAC). An ideal IAC network consists of nodes that can take on continuous values between a minimum and a maximum. The activation of the units is also supposed to change only gradually in time. This ideal is approximated by dividing time into a series of small steps. If we choose an activation function that cannot change very rapidly this discrete model acts as a good approximization for the ideal IAC-network.

The network (Figure (2)) consists of a set of nodes that are connected with links which can be excitatory or inhibitory (with a negative weight value). Some units can receive external stimuli, e.g. input from the syntactic module. The internal structure of a unit is shown in Figure (3). The input links are connected to a site that corresponds to their type. So each unit has distinct sites for external, excitatory and inhibitory links. The gate unit also offers a separate *gate* site with a special site function.

The site functions for the external, excitatory and inhibitory links simply compute the weighted sum of the input values *Iv*.

$$Sv = \sum_{i=1}^{n} w_i Iv_i$$

The site function for the gate site is a kind of "weighted AND" function. Its behaviour is similar to the weighted sum function when all input links have a value different from zero. However if one of the input links connected to the *gate* site is zero, the output *Sv* of the gate site function is also zero. The output of each site is scaled in order to control the influence of the different sites on the activation value.

$$Netinput = Sc_{inh}Sv_{inh} + Sc_{exc}Sv_{exc}$$

$$+Sc_{ext}Sv_{ext} + Sc_{gate}Sv_{gate}$$

The activation value *Av* for a new timestamp *t* can now be computed:
When *Netinput* is larger than zero:

$$Av^t = Av^{t-1} + (max - Av^{t-1})Netinput$$

$$-decay(Av^{t-1} - rest)$$

When *Netinput* is less than zero:

$$Av^t = Av^{t-1} + (Av^{t-1} - min)Netinput$$

$$-decay(Av^{t-1} - rest)$$

We see that the influence of *Netinput* on *Av* decreases when *Av* reaches its minimum or maximum value. On the other hand the influence of the decay rate is high in the upper and lower regions. When *Netinput* becomes zero, the Activation value slowly decreases to its rest value. The output value of the unit is equal to its activation, but only if the activation level is above a predefined treshold value. Otherwise the output is zero. So a unit with maximum activation that does not receive input anymore, slowly decreases its output value and than suddenly drops to zero beacuse its activation is below treshold value. This non linear behaviour is an essential property of connectionist models.

The bottom-up links are stronger than the top-down links because a unit may only be activated by bottom-up evidence. Top-down information may however influence the decision process at a lower level.

The values of the parameters in the model are:

Sc_{inh}	0.6
Sc_{exc}	0.6
Sc_{ext}	0.6
Sc_{gate}	0.6
treshold	0.5
decay	0.1
bottom-up weights	0.8
top-down weights	0.25
inhibitory weights	-0.8
external input weights	1.0
max	1.0
min	-1.0
rest	0

A simulation consists of a number of cycles in which activation spreads through the network. In each cycle the output and activation values for a time *t* are calculated from the values on time *t-1*. Figure (4) and (5) show the activation levels of the active units in the model: only activation levels above treshold (500) are displayed. At the beginning of the simulation all units are in rest state. We start the simulation for the disambiguation of "kick (the) bucket" by setting the output value of the external unit "kick" representing the output of a sub wordform level to 1. After three update cycles, the output of the external unit *II* (representing the fact that bucket is recognized) is set to 1. The duration of an external input is always one cycle. The availability of syntactic information is simulated by activating III_b and III_c before cycle seven. Figure (4) shows that the unit representing kick as a verb immediately follows this syntactic information and "kick as a noun" falls beneath activation treshold. After some more cycles a stable situation is reached (Figure (5)) which represents the best fitting hypothesis: the idiomatic reading.