Categorisation of Actions by Analogy:
from the Analysis of
Metaphoric Utterances to a Computational Model

Karine Duvignau
(duvignau@univ-tlse2.fr)
Lab. J. Lordat, UTM, 5 allées A.Machado
F-31058 Toulouse Cedex 1, France

Bruno Gaume
(gaume@irit.fr)
IRIT-UPS, 118 rte de Narbonne,
F-31062, Toulouse Cedex 4, France

Olivier Gasquet
(gasquet@irit.fr)
IRIT-UPS, 118 rte de Narbonne,
F-31062, Toulouse Cedex 4, France

Marie-Dominique Gineste
(gineste@lshs.univ-paris13.fr)
Lab. de Psychologie, Paris XIII, 99, av J-B Clément
F-93430, Villetaneuse, France

Abstract

If work in psychology has clearly brought to light the existence of "*conceptual flexibility*" in the categorisation of objects, which led to a re-questioning of the traditional conception of categorisation which considers rigid and discontinuous categories, it is not the case in linguistics and psycholinguistics. We propose, through a highlighting of the role of analogy in the categorisation of verbs, to defend the idea of *semantic flexibility* which constitutes a linguistic counterpart to psychologists' advances on categorisation. Accordingly, we show that the production of 'metaphoric' verbal utterances by adults and particularly by 2/3 year old children, reflects the existence of an *analogical categorisation of verbs* which makes it possible to argue in favour of a *computational model of the role of analogy* in the semantic network of the verb lexicon.

" *To say what a thing is, is to say what it is like*"
(Jackson 1866)

1. Introduction

Currently, the concept of "approximate identifications" proposed by Jakobson (1956) which suggests the existence of 'semantic flexibility' in linguistic categories, has not yet been truly validated. Thus, the metaphor, a linguistic phenomenon which consists in bringing closer distinct entities by substituting one for the other, is still considered primarily as a deviance while at the same time it could be the linguistic guarantee of the relevance of the categorial flexibility phenomenon brought to light by psychologists. This lexical connection by the metaphor which reveals a relation of semantic similarity, indeed highlights the "non-rigidity", the "pliability" of meaning of linguistic entities, properties which, in psychology, apply to concepts (Hofstadter & FARG 95). Our objective is to bring to light a continuum of meaning of linguistic entities starting from a highlighting of an **analogical** categorisation of verbs, then to propose a computational model of the role of analogy in the organisation of this continuum of meaning. For this purpose, we restrict our study of the verb lexicon,

a field still under-exploited, in particular for the French language, and we adopt two points of view: a linguistic and psycholinguistic approach of 'metaphoric' verbal utterances produced by adults and 2/3 year old children - key period in cognitive development (Bassano 2000)- and a mathematical analysis of the semantic network 'weaved' in language dictionaries by the verb lexicon.

The question of categorisation of actions is much less advanced than that concerning objects on which much work has been and is still devoted (Tomasello, Merriman 1995). Our methodological hypothesis for studying verbs on a psycholinguistic plane is that the 'metaphoric' utterances that adults produce and, more still, those which are uttered by 2/3 year old children, who learn the meaning of the words at the same time as they build their categories (Schlesinger 1982), reveal the method of categorisation of actions.

Certainly, that the utterance of one verb could be 'metaphoric' w.r.t. another one (an 'exact' or more precise one) should have something to do with semantic similarity between those verbs. According to (Resnik and Diab, 2000) there are mainly three approaches that addresses semantic similarity between words depending on the kind of models they rely on: a) taxonomic models (TM) –à la WORDNET (Fellbaum, 1998)— like those presented for example in (Resnik, 1999, Lin 1998), b) distributional co-occurrence model (DCOM) of Schütze (1993) for example, and c) semantic structure models based on the theory of lexical conceptual structure (LCS) of Dorr (1993), Jackendoff (1983). These approaches provide semantic similarities that range from the more purely semantic (with TM more sensitive to semantic content of words), to the more syntactic (with LCS more sensitive to the semantic structure of words), DCOM being sensitive to both semantic and syntactic features. Not each of them are presented as cognitive models but we follow Resnik and Diab in saying that comparisons with human ratings of similarity should provide a useful base for comparing similarity measures. First, such comparisons with human judgements have mainly involved nouns, hence only few data are available concerning verbs. Still, the work

presented in (Resnik and Diab, 2000) proposes a methodology for such comparisons and analyses first results: globally each approach perfoms well in capturing aspects relevant w.r.t. it is based on, but they all fail in predicting similarity between verbs rated low but non zero by humans. They cite some examples like that of the pair "unfold/divorce": these verbs are not related in WORDNET taxonomy, they have little semantic structure in common and do not sufficiently co-occur in corpuses to be statiscally related. From these examples they suggest that these approaches fail for pairs of verbs which are in metaphorical relation. Thus, today, no computational approach connects the verbs/actions « éplucher» (*to peel)* and « déshabiller » (*to undress)* which refer to distinct categories, while at the same time their connection arises both in children's productions as of 2/3 years of age - e.g. « je déshabille l'orange» (*I undress the orange*), and in those of adults - e.g. « ... raccommoder les principes...? » ("*... to darn the ... principles ?*") (cf. sect. 2, Corpus, ex. 1 and 3). To qualify these utterances, we introduce the concept of *approximation by analogy* (developed in section 2) which makes it possible to underline the role of analogy in the categorisation of verbs/actions.

The corpus of 'metaphoric' productions by children and adults that we have collected up to now supports the psycholinguistic hypothesis whereby the mental lexicon is organized in a network around metaphoric poles (Jakobson, 1956).

Our objective is to validate this organisation of the lexicon, and in addition to propose an associative computational model of it. For this reason we postulate the cognitive hypothesis whereby traces of this network organisation are present in the language dictionaries which from their method of constitution provide 'naturally' a reflection of the lexicographers' mental lexicons.

2. *Semantic Approximation by Analogy* and Categorisation of Verbs/Actions

In the field of early lexicon acquisition, the advances on the nature of the lexiconising processes relate primarily to the nominal productions of children: the works (Anglin 1976; Clark 1993; Dromi 1999) establish that the child, wanting to name an entity for which he does not have a substantive available, would use analogy, i.e. a form of resemblance between properties of a known and lexiconised object and those of another new object which needs to be indicated. It is only recently that attention has moved to the categorisation processes of verbs (Tomasello & Merriman 1995) and one can legitimately raise the question: according to which methods their categorisation takes place. To this end we make the following hypothesis:

The child seeking to communicate an event A [to tear a book] for which he does not have a constituted verbal category: 1) would make an analogy with an old event B [to break a glass] already memorised with a lexical entry

"to break" and 2) using this analogy, would say "the book is broken" to communicate event A.

Depending on the case, he will produce (a) hyperonymic or synonymic, or (b) 'metaphoric' expressions which will have the effect of making up for his lexicon deficit. (cf Corpus 2). To function by analogy in the production of verbs supposes that the child has already worked out elementary semantic relations (agent-action, action-patient; agent-action-patient; agent-action-localisation). These semantic relations are assumed to be acquired as soon as the young child is able to combine two or three words, i.e. as from 20-24 months of age: even if the child does not use the canonical forms (by marking his utterances syntactically) he produces ordered series which express these different semantic relations (Brown 1973). But it is not enough to have worked out these semantic relations. The child, to be able to use the known verb and to apply it to a new situation, must be able to discover a relation of analogy between two events. We consider that the young child, when he speaks, is in fact in a situation of having to make himself understood by using adequate utterances. Yet his lexicon being extremely limited at 2 years of age and still restricted at 3.5, he will have then to draw from this lexicon and to use items not randomly, - communication then being likely to fail -, but according to a principle of *relevant analogy*.

Corpus

We examine the verbal productions of 2/3 year old children (Corpus 1) to highlight this capacity.

With regard to the metaphoric productions of adults (Corpus 2), we consider that they express their capacity to connect an entity which concerns a semantic field A with an entity which concerns a semantic field B. Our hypothesis is that these metaphoric productions reflect an implicit categorisation of verbs by an analogical, sometimes creative, mode, (Gineste, Indurkhya, Scart 2000) and they enable to confer on the metaphor the status of *intralingual translator* (Duvignau 2001) which consists in transmitting, from one semantic field to another, a concept that these two fields have in common.

Corpus 1 (131): Spontaneaous Utterances Collected From French-Speaking Children Aged 24 To 36 Months. (Duvignau 2001) Extract:

1- « je déshabille l'orange » 36 mois, [éplucher…] (l'enfant épluche une orange) / "I undress the orange" 36 months, [to peel…] (the child peels an orange)

2- « maman, tu peux coller les boutons » 36 mois, [coudre...] (les boutons sont décousus, il faut les coudre) / "mum, you can stick the buttons" 36 months, [to sew...] (the buttons are unstitched, they have to be sewn on)

Corpus 2 (400): Utterances Collected From Scientific Texts With Didactic Aims. (Duvignau 2001) Extract:

Henri Poincaré (1905) « La valeur de la science »

3- « Faudra-t-il chercher à raccommoder les principes ébréchés ? » [réviser...] / "Will it be necessary to try to darn the chipped principles ?" [to revise...]

4- « l'analyse ne peut séparer deux faits simultanés sans les mutiler ». [altérer...] / "analysis cannot separate two simultaneous facts without mutilating them". [to alter...]

We consider these productions by adults and children as *semantic approximations by analogy*, which enables the abandonment of the 'error' or even 'metaphor' concept, within the framework of children's productions.

3. Computational *Metaphorymy* Model

The computational model that we develop proposes to account for the categorisation of verbs as well as their analogical relation by establishing similarity between verbs. Contrary to feature-based similarity, this relation of similarity is not established by 'local' comparison of the characteristics of two given verbs, but by considering their respective 'global' positions in a network.

Generally, if dictionary definitions carry meaning, it is at least by the network that they establish between the words which are the entries. The idea of exploiting this network (regarded simply as a structured textual source) was implemented by Ide and Véronis (1990) through a network of neurons for disambiguation[1], but this work did not meet the expectations of its authors. It was also exploited in (Ploux and Victorri 1998) starting from dictionaries of synonyms with a view to building semantic spaces. We formulate the psycholinguistic hypothesis that the network the dictionary definitions 'weave' is a trace of the mental lexicon distributed around metaphoric poles, this guides the mathematical analysis that we make. We analyse this network through its representations in the form of graphs.

We have developed an algorithm **MET** which, starting from a dictionary graph makes it possible to calculate similarities between its nodes (the entries). Example: list of nodes similar to the verb ÉCORCER, (*TO BARK*[2]) (most similar to least similar) calculated on the verbal entries of our dictionary[3]:

[ÉCORCER, DEPOUILLER DEPIAUTER, PELER, DECORTIQUER, ÉPLUCHER, ÉCORCHER, ENLEVER, OTER,... DÉNUDER, PLUMER, EFFEUILLER..., DÉSHABILLER, DÉVÊTIR..., ANALYSER, EXAMINER, ...]

[*To Bark, To Strip, To Skin, To Peel, To Shell / To Hull, To Peel, To Skin / To Graze, To Remove , To Take Off... To Bare / To Strip, To Pluck, To Thin Out the leaves of..., To Undress / To Strip, To Undress..., To Analyse, To Examine...*]

One can notice that the verb DEPOUILLER (*TO STRIP*), which appears at the head of the list (most similar to ÉCORCER (*TO BARK*) according to MET) is in

hyperonymic relationship to the verb ÉCORCER (*TO BARK*).

The similarity[4] calculated by the algorithm MET organises in a continuum the three linguistic concepts of **hyperonym** and **synonym** (by the most similar nodes) and that of **metaphor** (by the nodes a little less similar). The introduction of the concept of *metaphorymy* which covers these three concepts makes it possible to underline the continuous slip of meaning that exists from a word in synonymic relation, to a word in metaphoric relation, as the similarity with the word of reference decreases. Reinvesting the concept of distance between lexical fields (Tourangeau and Sternberg 1982, Barsalou 1989), makes it possible, in particular, to refine the lexical relation of synonymy and to reconsider the concept of metaphor. Indeed, if one examines for example the similarity obtained by MET between a target like "bark" and another term:

- when the similarity is strong, as between « écorcer » (*to bark*) and « éplucher » (*to peel*), the terms are bound by a relation of **"intra field" similarity**, which connects terms concerned by the same semantic field (in this case /vegetable/) while

- when the similarity is a little less strong, as between « écorcer » (*to bark*) and « déshabiller » (*to undress*), the terms are bound by a relation of **"inter field" similarity** which relates to terms concerned with different semantic fields (in this case /vegetable/ and /body/).

This "inter field" similarity constitutes a new conception of the metaphor considering, as Tourangeau and Sternberg (1982) underline within their research framework, the greater the distance is between two terms, the sharper the metaphor can be.

This mathematical approach of the lexicon highlights a distribution of verbs around **metaphorymic poles**, in echo of Jakobson's (1956) "metaphoric pole". It suggests allotting to the concept of metaphorymy the status of lexical relation and contributes to validating the concept of "approximate identifications" introduced by Jakobson to indicate the fact of "To say what a thing is, is to say what it is like" (Jackson 1866).

---

[1] Recognition of the meaning of a word among those given by example in a dictionary, or distinction of a word among its different homographs.

[2] 'To bark' is the least bad translation of the french 'écorcer' (to strip of its bark a tree).

[3] Obtained from a compilation of several dictionaries.

[4] We qualify as **similarity** the binary relation **MET-G(X,Y)** that MET calculates on S_G the nodes of G, as for any graph G the properties of MET-G are:

1) $\forall X \in S_G$ MET-G(X,X) = 1 and $\forall Y$, MET-G(X,Y) $\in$ [0,1]

2) In general MET-G(A,B)≠MET-G(B,A), that is MET-G(_,_) is not a symmetric relation. For example:

MET-G(DEPOUILLER,ECORCER) >

MET-G(ECORCER,DEPOUILLER) but however if MET-G(A,B) is large relative to all values {MET-G(A,X)/X $\in$ S_G} so MET-G(B,A) remains quite large relative to all the values {MET-G(B,X)/X $\in$ S_G}.

As was underlined in (Tversky 1977) and was recalled in (Love 2000), asymmetry is necessary for models of similarity.

Methodology

All dictionaries can be represented by graphs whose nodes and edges can be defined in multiple ways. One of which consists in taking for nodes of the graph the dictionary entries and to admit the existence of an edge of a node A to a node B if and only if the entry B appears in the definition of entry A. This is the starting position which we have adopted. Indeed, this procedure makes it possible to extract from any dictionary, which henceforth we will call the **graph of the dictionary**.

Illustration: Definition of ÉCORCER (*TO BARK*)

ÉCORCER [ekóRse] v. tr.; Dépouiller de son écorce (un arbre). Décortiquer, peler (le grain, les fruits)

*(**TO BARK** [ba:k] v. tr.; To strip of its bark (a tree). To shell, to peel (grain, fruits))*

Starting from this definition, and by applying our procedure (the arrow thus means "has in its definition"), one obtains:
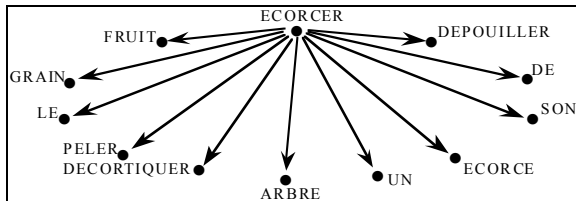

Figure 1. Extract of the graph around ÉCORCER (*TO BARK*)

By reiterating this construction for each entry of the dictionary, the **graph of this dictionary** is obtained. If one is only interested in the verbs which appear in the infinitive in the entries which are verbs here is what we obtain "around" the node indicated by the verb ÉCORCER (*TO BARK*):
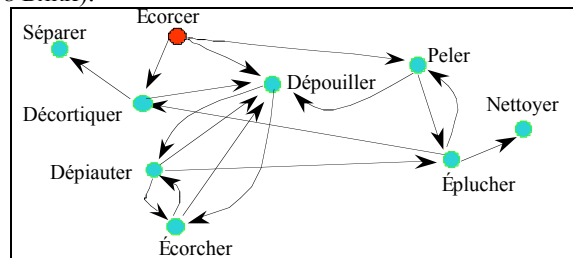

Figure 2. Verbs graph around ÉCORCER (extract)

The definitions NETTOYER (*TO CLEAN*), SEPARER, (*TO SEPARATE*)... refer to other verbs absent from our diagram for reasons of legibility (if one continues, one would quickly come across every verb in the dictionary !). Thus we transferred to this figure only part of the neighbors of order 1, 2 and 3 of ÉCORCER (*TO BARK*). Once this graph is obtained, our algorithms work from what we call an anonymous graph.
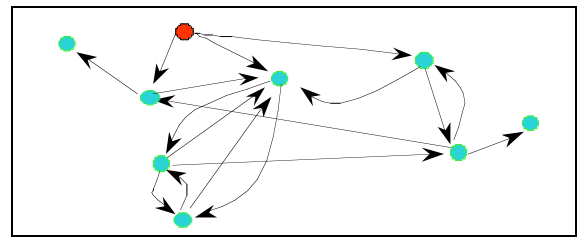

Figure 3. Extract of the anonymous verbs graph, around the associated node of ÉCORCER (*TO BARK*)

Our hypothesis is that in the very structure of a dictionary graph (its anonymous graph) information of a semantic nature remains, encoding the semantic similarity. To reveal the structure of these graphs, we develop methods which associate the combinative and the statistical. This type of method has already proved its effectiveness in fields as varied as Data-Mining it (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy 1996), the pagination of electrical supply networks (De Frayssaix, Kuntz 1992), the analysis of the visual cortex structure (Jouve, Rosenstiehl, Imbert 1998) or the structure of the Internet network (Hassenforder, Ferré, Jouve 2000). The common point of these works is, starting only from an examination of the structure of the anonymous graph obtained, to manage to extract the **topologico-semantic** properties. The search for these properties is done in several stages: the first carried out by MET[5] consists of a calculation of similarity between the nodes of the graph. The important idea is to calculate the similarity between two nodes starting from the globality of the graph. That means that, contrary to the methods quoted above, the immediate neighbors of two nodes are not only taken into account for the calculation of their similarity, but the totality of the graph (holistic approach). The second stage is the extraction of the graph's structure starting from the table of similarities, by a traditional method of automatic classification of this table. It is by applying this method of analysis to language dictionaries that we highlight the structure of their graphs and 'capture' their topologico-semantic properties among which appear metaphorymy which organizes in a continuum the **hyperonymy-hyponymy, and the similarity intra and inter field**. Let us note that contrary to all similar approaches, we have the possibility of extracting the hyperonymic relation by not taking

---

[5] MET is a rather simple algorithm, but which it would be too long to develop formally here. The main idea is that MET calculates primarily a **swarm** concept or 'social clique' (zone with strong density of links): a word A is close to a word B if word A has many links with words which link with B (the friends of my friends are my friends and the more friends we have in common, the more we are friends). The concept of swarm is less constraining than the concept of clique. A node can be very similar to another without there being an edge between them (it is the case for example of ÉCORCER (*TO BARK*) and DÉPIAUTER (*TO SKIN*) which are found to be very similar although they do not have any direct link in the Grand Robert).

absolutely into account the position of the verbs in the definition (let us not forget that MET only works on the anonymous graph).

4. Conclusion

We are currently carrying out a longitudinal experiment[6] on a troop of 60 children (2 to 4 years of age). We measure on which processes are based the analogies present in their semantic approximations, as well as their dynamics. We will then be able to compare the semantic similarities observed in children with the mathematical similarity calculated by MET within a longitudinal experimental framework.

Currently, from the results of our study of the two corpora that we have collected (adult and children) we can confirm the role of analogy in the categorisation of verbs/actions and propose to allot to the concept of *metaphorymy* the status of lexical relation, which offers a new conceptual framework in linguistics for the establishment of a categorisation of verbs. The computational model MET allows the automation of this metaphorymic classification of the lexicon from the graphs of dictionaries and seems well able to be used as a basis for the development of a continuous model of the meaning of lexical entities of a language: it may help in filling the gap left open by other models w.r.t. metaphorical relation between words. This model is based on the primitive concept of 'link', (no features, nor prototypes, nor even exemplars) from which MET builds *swarms* which reveal verbal categories of which analogy is the fundamental cognitive ergonomic principle. The **analogic link** being the 'pixel' from which MET draws the categorial image of the meaning of the lexical entities present in the graphs of dictionaries.

References

Anglin JM (1976), Les premiers termes de référence de l'enfant. *Bulletin de Psychologie*, 232-241.

Barsalou L. W. (1989) Intraconcept similarity and its implications for interconcept similarity. In S. Vosnadiou & A. Orthony (eds), *Similarity and analogical reasoning*. Cambridge University Press.

Bassano D. (2000) Early development of nouns and verbs in french; exploring the interface between lexicon and grammar, *Journal of Child Language*, 27: 521-559.

Brown R. (1973). *A first language: the early stages*. Cambridge, CA: Harvard University Press.

Brooke-Rose C. (1958) *A grammar of metaphor*, Seeker and Warburg

Clark, E.V. (1993) *The lexicon in acquisition*. Cambridge University Press

De Frayssaix H., Kuntz P. (1992) Pagination of large-scale networks, *Algorithms Review*, 2:105-112.

Dorr, B.J. (1993) *Machine translation: a view from the lexicon*. The MIT Press ,Cambridge, MA.

Dromi E. (1999) Early lexical development. In Barrett (Ed.), *The development of language*, Psychology Press.

Duvignau K. (2001) De la métaphore analogique à l'acquisition précoce du lexique: l'analogie pour l'apprentissage des verbes. *Proceedings of Early Lexicon Acquisition*, ELA 2001, Lyon, France, (à paraître).

Fayyad U.M., Piatetsky-Shapiro, G., Smyth, P & Uthurusamy, R. (Eds.) (1996) From data mining to knowledge discovery: an overview, in *Advances in Knowledge Discovery and Data Mining*, CA: AAAI Press.

Fellbaum C. (1998) A Semantic Network of English Verbs. In Fellbaum C. (Ed.) *WordNet*.

Gineste M-D., Indurkhya B., Scart V. (2000) Emergences features in metaphors comprehension, *Metaphor and Symbol*, 15(3):117-135.

Hassenforder C., Ferré L., Jouve B. (2000) Partitioning of directed networks: the WWW case, *2nd International Workshop Network Sampling*,The Nederlands.

Hofstadter D. and the Fluid Analogies Research Group (1995) *Fluid concepts and creative analogies*. New york: Basic Books.

Ide N., Véronis J. (1990). Word Sense Disambiguation with very large neural networks extracted from machine readable dictionaries. *Proceedings of the 14th International Conference on Computational Linguistics*, Helsinki.

Jackendoff, R. (1983) *Semantics and cognition*. The MIT Press, Cambridge, MA.

Jackson H. (1866) Notes on the physiology and pathology of language, *Brain*, 1915, XXXVIII,48:58.

Jakobson R., (1956) Two Aspects of Language and two Types of Aphasic Disturbances, In Jakobson R., Halle M., Fundamentals of Language, Mouton & Co.

Jouve B., Rosenstiehl P., Imbert M. (1998) A mathematical approach to the connectivity between the cortical visual areas of the macaque monkey. *Cerebral Cortex*, 8:28-39

Lin D. (1998) An information-theoretic definition of similarity. In *Proc. of the 15th Int. Conf. on Machine Learning (ICML'98)*, Madison, Wisconsin.

Love, B. C. (2000) A Computational Level Theory of Similarity. *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, Erlbaum Associates.

Ploux S., Victorri B. (1998) Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39(1):161-182.

Resnik P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95-130.

Resnik P., Diab M.(2000) Measuring Verb Similarity. *22nd Annual Meeting of the Cognitive Science Society (COGSCI2000)*, Philadelphia, August 2000.

Schlesinger, I. M . (1982). *Steps to Language: Toward a Theory of Language Acquisition*. Hillsdale, Lawrence Erlbaum associates

Schütze, H. (1993) Word space. In Hanson, Cowan, Giles (eds), *Advances in neural information processing systems* 5, Morgan Kaufmann Publishers, CA.

Tomasello M. & Merriman W. E., (Eds).(1995) *Beyond names for things*, Hillsdale, NJ: L. Erlbaum Associates.

Tourangeau R., Sternberg R. (1982) Understanding and appreciating metaphors, *Cognition* 11:203-244

Tversky A.(1977) Features of similarity. *Psychological Review*, 84: 327-352.