

# ANALYSING DICTIONARY DEFINITIONS OF MOTION VERBS

ANTONIETTA ALONGE

Università di Pisa - Dipartimento di Linguistica  
Via S. Maria, 36 - 56100 PISA - Italy  
Tel. +39-50-560481; Fax +39-50-589055  
e-mail : LEMINTER@ICNUCEVM.

## 1. Introduction

In recent years many researchers working in the field of NLP have turned to large text resources such as machine-readable dictionaries (MRDs) and text corpora, in order to cope with the major problem of building a large computational lexicon. Within the Esprit project "Acquilex"<sup>1</sup> the possibility of exploiting existing mono- and bi-lingual MRDs of four languages<sup>2</sup> in order to develop a multi-lingual and maximally reusable (by different researchers, for different NLP tasks) lexical knowledge base (LKB) is being explored and, indeed, we are finding evidence that important semantic and syntactic information can be semi-automatically extracted from these sources with significant saving in resources, compared to building a lexicon by hand.

The task of extracting lexical information from MRDs is a non-trivial one in that such information is mostly implicit in dictionaries and completely new techniques and methodologies have to be developed in order to achieve this goal. Furthermore, it is necessary both to take into account the peculiar characteristics of dictionary structures and to make theoretical hypotheses about the kind of information which can be useful for NLP systems and which therefore should be extracted. Indeed, these are two of the major issues being dealt with within Acquilex (another one being the issue regarding the representation of the knowledge extracted). Therefore, our work has been guided by theoretical hypotheses and empirical observations at the same time. First of all we assumed the centrality of the lexicon in the organization of natural languages; then, on the basis of the growing interest of different theoretical frameworks for semantic phenomena and of the fact that contemporary syntactic theories seem to converge on the hypothesis that syntactic structure is, to a large extent, determined by word meaning, we tried to see if it was possible to identify, within our dictionaries, that kind of semantic information on verbs which had been described as determining fundamental syntactic behaviours of the verbs themselves (cf. Levin 1985; Levin & Rappaport 1991). Finally, we tried to follow some indications, provided in works such as those of Pustejovsky (1989) or Boguraev &

Pustejovsky (1990), relative to the kinds of lexical data which should be sought within MRDs and other computerized sources in order to be able to deal with problems facing the computational linguist aiming at building components for NLP. According to Boguraev & Pustejovsky (1990: 39), i.e., the following information should be individuated within sources of lexical data such as MRDs: argument structure; event structure (Aktionsart); qualia structure (Pustejovsky (1989)); lexical inheritance structure.

Dictionary definitions are generally structured so that it is possible to identify two fundamental parts within them: the "genus term", which is connected with the entry-word by means of an IS-A (or taxonomical) relation, and the "differentia" part, in which what differentiates the entry from its hypernym is indicated<sup>3</sup>. Within Acquilex we are improving and developing techniques aimed at exploiting both the taxonomical organization of information in dictionaries and the fact that recurrent structures and patterns are found in the differentia; such patterns are being individuated through a pattern-matching procedure (which operates on the output of the syntactic analysis of definitions) in order to associate them with corresponding relations or semantic / conceptual categories (for descriptions of methodologies for extracting taxonomies from MRDs, see Calzolari (1984) and Chodorow, Byrd & Heidorn (1985). For discussion of this way of analysing the differentia, see Calzolari (1991)). In the following, a research being carried out within this project and aimed at developing techniques for extracting different kinds of information related to motion verbs is presented.

## 2. Information on motion verbs within dictionaries

Motion verbs have been the subject of several studies by linguists because they present particularly interesting semantic and syntactic characteristics. In particular, even if they are often considered as being a coherent semantic class (and indeed we speak of "motion verbs" as a whole), we can find verbs displaying different semantic features and syntactic behaviour in this class. With our research,

therefore, we have tried to individuate information on this class of verbs to be used to further classify them according to semantic characteristics which subsets of them present and which are connected with different syntactic behaviours.

By analysing our mono-lingual Italian dictionaries and the Collins bi-lingual, it is possible to extract the following kinds of syntactic and semantic information on motion verbs, by means of semi-automatic procedures:

- 1) transitivity / intransitivity / reflexivity<sup>4</sup>;
- 2) Aktionsart (or "lexical aspect");
- 3) unaccusativity or unergativity (for intransitive verbs);
- 4) components of meaning, typical subjects, thematic roles.

While the information in the first point is explicitly coded within our dictionaries, a procedure was developed in order to extract information on Aktionsart (i.e., to classify verbs in a Vendlerian fashion (Vendler (1967))) which exploits the taxonomical organization of data in dictionaries and the possibility of having inheritance of information as a consequence of the IS-A link. After classifying "genus term" verbs, we make each hyponym inherit the Aktionsart-class from its superordinate verb unless some specific patterns, recognized through a pattern-matching procedure, are found in the differentia part of its definition. If this is the case, the entry-verb considered is classified in a different way, according to specific rules stated in advance (see Alonge (1991)).

Information on unaccusativity or unergativity of intransitive motion verbs (in Perlmutter's (1978) sense) is easily extracted for Italian by taking into consideration the auxiliary selected by a verb (unaccusatives take *essere* (to be)), which is indicated in the Collins bi-lingual.

Finally, by analyzing the differentia part of definitions, we find information on components of meaning of verbs, typical subjects and also thematic roles. Components of meaning and typical subjects are individuated by identifying recurrent patterns, clearly referring to specific semantic categories, within the differentia. For the time being, a detailed analysis has been carried out in relation to the verbs within the taxonomy of *muoversi* (intransitive to move) in GRZ (244 entries); we first examined manually some of the definitions of verbs within the taxonomy and individuated recurrent patterns connected with components of meaning which were, therefore, considered potentially relevant to describe the semantics of the whole class of verbs, even if not every pattern was found in each definition. The following are examples of the patterns found within the definitions analysed and of the components of meaning which were connected with them:

- MANNER of MOTION: con / come / a NP; AdvP; V-ing
- GOAL: a / incontro a / verso NP; AdvP
- SOURCE: da NP
- PATH: da ... a / da ... verso NP
- MEDIUM: per via di / in / a NP
- PURPOSE: per VP
- TYPICAL SUBJECT: detto di / si dice di / di NP

A pattern-matching procedure is now being implemented which identifies patterns within the differentia and relates them to semantic categories (or typical subjects). Furthermore, each verb inherits the components of meaning connected with its superordinate verb(s)<sup>5</sup>. *Agitarsi* (to toss) is defined (in GRZ, sense 1) as "muoversi con vivacità, con irrequietezza, con violenza" ("to move vivaciously, restlessly, violently"); the genus term *muoversi* simply refers to "the fact of motion" and is not connected with any other particular components of meaning; the PPs which we find in the differentia, instead, refer to MANNER of MOTION, so that our procedure will analyse *agitarsi* as connected with a MANNER of MOTION component of meaning. *Fuggire* (to escape), then, is defined (GRZ, 1) as "allontanarsi di corsa, per lo più per evitare un pericolo o un danno" ("to go away at a run, mostly to avoid a danger or a damage"); since the genus term *allontanarsi* is related to a GOAL component of meaning, *fuggire* itself will be connected with the same component of meaning plus a MANNER of MOTION component (di corsa) and a PURPOSE component (per evitare ...).

Sometimes different semantic categories may be indicated by the same lexical category / pattern, so that it was necessary to define lists of specific (sequences of) words to be connected with one of the components of meaning in order to distinguish instances of it from instances of the other component related to the same pattern. For instance, the same preposition can be used in Italian to express different meaning categories. This is the case of the preposition *a* (at, in, to...): when coupled with most nouns (/ noun phrases) it indicates GOAL (with motion verbs); however, when it is used in conjunction with certain nouns, idiomatic expressions are formed which refer to MANNER of MOTION. Therefore, in order to recognize the latter cases, we identified a limited list of such idiomatic expressions (which can be found with motion verbs) so that when one of these expressions is found within a definition the verb defined will have a MANNER of MOTION component of meaning, otherwise, it will have a GOAL component<sup>6</sup>. *Andare* (to go), e. g., is found together with "a cavallo" (riding) within the definition of *cavalcare* (GRZ, 1) (to ride)

and it is also found with "a letto" (to bed) within the definition of *coricarsi* (GRZ, 1) (to go to bed). In the first definition the idiomatic expression "a cavallo" indicates MANNER of MOTION, while in the second definition we have the indication of a GOAL; therefore, *cavalcare* will be connected with a MANNER of MOTION component, while *coricarsi* will refer to motion towards a GOAL. Similar procedures are being applied to other cases of ambiguities and it seems possible to claim that, even if it is necessary to do some work by hand, the utilization of MRDs and of such semi-automatic methodologies of analysis represents a significant saving both in time and resources.

The same information on components of meaning was also utilized to identify thematic proto-roles, according to Dowty's (1988) proposal, which has been adopted within Acquilex (Sanfilippo (1991)). Dowty individuated two sets of properties which contribute to the definition of "prototypical" agent and patient role and which are entailed by verb meaning:

• CONTRIBUTING PROPERTIES FOR THE PROTO-AGENT ROLE:

volition, sentience (and / or perception) causes event, movement

• CONTRIBUTING PROPERTIES FOR THE PROTO-PATIENT ROLE:

change of state, incremental theme, causally affected by event, stationary.

The argument having the highest number of proto-agent properties entailed by the meaning of the verb, and inherited by default, is to be associated with the proto-agent role; the argument of a transitive verb to which the highest number of proto-patient properties can be ascribed (inherently via entailment relations, and by default), instead, is to be associated with the proto-patient role (cf. Sanfilippo (1991)).

Thus, since fundamental questions about the identification, individuation, and even the theoretical status of "traditional" thematic roles remain unresolved, we decided to determine the semantic content of these basic roles by taking into account properties which are needed for verb classification and which can be identified through the analysis of definitions described above.

By examining the verbs within the taxonomy of *muoversi*, we saw that even if they can be either strict intransitives, or strict transitives, or intransitives taking an oblique object, they all imply a subject argument (a proto-agent) which corresponds to the "moving object" and for which either the manner of motion or the direction (and therefore, a change of position) can be inherently specified. The information that it is the subject of these verbs which is moving is inherited from the genus term *muoversi*; the

specification relative to the manner of motion or the change of position is found within the differentia of definitions and used to encode more information in relation to the "moving object" itself. I.e., if we take into consideration the definitions of the verbs *andare* (to go) and *oscillare* (to swing) given below, we may see that in relation to the former verb the proto-agent moves along a path, while in relation to the latter the manner of motion of the proto-agent is inherently specified:

• *andare* : muoversi da un luogo verso un altro (GRZ, 1) (to move from one place to another)

• *oscillare* : muoversi alternamente in qua e in là o in su e in giù (GRZ, 1) (to move alternately here and there or up and down)

### 3. Data from dictionaries and theoretical works

The data we have extracted from dictionaries were indicated as necessary for a computational lexicon within theoretical works (cf. above). Furthermore, the data related to components of meaning can be compared to the results of theoretical works in which connections among semantic and syntactic characteristics of motion verbs were identified, in order both to verify the validity of hypothesis put forward on the basis of observation of limited data and to derive information on syntactic behaviours of large amounts of verbs.

Talmy (1985) deeply investigated the relationships among surface expressions related to motion verbs and their semantics across languages. He individuated different ways of "conflating" components of meaning across languages which relate to the different syntactic configurations allowed. The ability to refer to these lexicalization patterns, some of which are universal across languages while some vary systematically defining typologies of languages, can help structure a multi-lingual LKB of the sort we are building. Actually, in dictionary definitions we find useful information in relation to the kind of "conflation" of components of meaning which are allowed in Italian and it is possible to compare our data with Talmy's analyses. While English (but also Chinese, etc.) can express at once the "fact of Motion" and its manner, (post-Latin) Romance languages cannot, according to Talmy. However in Italian many verbs express at once "the fact of MOTION" and MANNER; what they do not often express is what Talmy (1985: 141) calls "translational motion", i.e. change of position, and MANNER together<sup>7</sup>. For most verbs in Italian what is relevant is either the manner of motion, like, e.g., for *camminare* (to walk) or the change of position, like, e.g., for *coricarsi* (to go to bed). Actually, in dictionary definitions

we find information on this kind of “conflation” of components of meaning:

• *camminare*: andare a piedi (GRZ, 1) (the pattern underlined refers to MANNER of MOTION);

• *coricarsi*: andare a letto (GRZ, 1) (the pattern underlined indicates GOAL).

Nevertheless, in Italian there are some manner of motion verbs which behave somewhat differently. They are manner of motion verbs which are both unergative and unaccusative. *Correre* (to run) has both an unergative and an unaccusative use; when it used in the unaccusative form it refers both to MANNER of MOTION and GOAL, as can be seen in the examples below:

• Giovanni ha corso per tre ore / \*a casa

• Giovanni è corso a casa

Both sentences can indeed be translated as “Giovanni ran”, but only with the unaccusative form (with the auxiliary *essere*), the goal expression “a casa” is allowed. Information on this characteristic of *correre* can be found within DMI, where even if *correre* is defined as “andare velocemente” (“to go fast”, where the adverb refers to MANNER of MOTION), it is also stated that when this verb is used with the auxiliary *essere* (and, therefore, is unaccusative) it implies a GOAL.

Italian may express at once, as noted above, the fact of motion and a GOAL, but also motion plus SOURCE or PATH, with such verbs like *entrare*, *uscire*, *passare*, *salire*, etc., which have direct counterparts in English, even if “these verbs (and the sentence patterns they call for) are not the most characteristic of English”, according to Talmy, (1985: 72); indeed verbs such as *enter*, *exit*, *pass*, *descend*, etc. are borrowings from Romance. The fact that this kind of conflation is typical of Italian (and Romance) but not of English is further demonstrated by the existence of verbs such as the above mentioned *coricarsi*, or *esulare* (to go into exile), etc. which have no direct correspondent verbs in English. Also with respect to these verbs we find useful information within dictionary definitions, as can be seen in the examples below:

• *entrare*: andare dentro (GRZ, 1)

• *salire*: andare verso l'alto, in su (GRZ, 1)

• *uscire*: andare fuori (GRZ, 1)

(patterns indicating PATH / GOAL have been underlined).

According to Talmy, then, motion can never be conflated with PURPOSE. Nevertheless, among the verbs we analysed there are some which seem to incorporate a purpose together with motion. *Passare* is defined (in GRZ, 1) as: “muoversi attraversando, percorrendo un luogo, per andare in un altro” (“to move crossing one place, in order to go to another one”), where

“per andare in” refers to the purpose of the action indicated by the verb. If *passare* has to be seen as having this component of meaning, then all its hyponyms should inherit it and so there would be some verbs indicating at once motion and purpose<sup>8</sup>.

Levin and Rappaport (1991) further investigated intransitive motion verbs and claimed that, on the basis of their syntactic behaviour / semantic features, it was possible to distinguish among three classes of intransitive motion verbs:

1) **arrive** verbs (unaccusative, implying a component of meaning DIRECTION and telic);

2) **run** verbs (unergative; components of meaning MANNER of MOTION + PROTAGONIST CONTROL / NO DIRECT EXTERNAL CAUSE);

3) **roll** verbs (unaccusative; components of meaning MANNER of MOTION + NO PROTAGONIST CONTROL / DIRECT EXTERNAL CAUSE).

Furthermore, these classes can be related systematically to Vendlerian classification (based on *Aktionsart* distinctions; Vendler (1967)).

Information found within intransitive motion verb definitions, therefore, was used also for dividing verbs according to the distinctions individuated by Levin & Rappaport.

By analysing our data we found evidence that the component of meaning which is relevant to identify “arrive” verbs is that of GOAL (DIRECTION); furthermore, the component of GOAL seems to be relevant also when it is missing. That is, the lack of such a component indicates manner of motion, even if we do not find patterns related to MANNER within the definition. *Volare* (GRZ, 1) is defined as “muoversi in aria” (“to move in the air”), where “in aria” indicates the MEDIUM and not the manner of motion. However, the fact that we do not find an indication of a GOAL component seems sufficient for classifying the verb as a manner of motion verb and not a change of position one. In order to decide, then, if it is a “run” or a “roll” verb (see above) we use the information on unaccusativity / unergativity, since in our dictionaries we do not find any references to the existence of control on the part of an agent. Actually, further study seems to be needed with respect to such a component of meaning and its relation to unaccusativity / unergativity, because by analysing our data we found manner of motion unaccusative verbs which seem to imply protagonist control (which would contradict the hypothesis put forward by Levin and Rappaport).

#### 4. Conclusion

In this paper it has been shown how, by

combining theoretical assumptions with empirical observations, it is possible to individuate information to be used for building a computational lexicon for NLP within dictionaries. At the moment, we are working both for automating the various stages of the analysis and for encoding the information extracted in the multi-lingual LKB which we are building (which uses a typed system of unification as representation language). At the same time, our data should be related to data from the other dictionaries used within the project.

<sup>1</sup> Esprit - BRA 3030 on the "Acquisition of Lexical Knowledge for NLP". Research teams from the Universities of Amsterdam, Barcelona, Cambridge, Dublin and Pisa are collaborating within this project.

<sup>2</sup> In Pisa two dictionaries of Italian, are being used: the Nuovo Dizionario Garzanti (GRZ) and the Dizionario-Macchina dell'Italiano (DMI), a MRD mainly based on the "Vocabolario della lingua Italiana" Zingarelli. Furthermore, we also utilize the Collins bi-lingual Italian-English dictionary.

<sup>3</sup> As a matter of fact, not all definitions are structured in this way. Indeed, other relations (like, e. g., PART-WHOLE relation) are found. However, as far as verb definitions are concerned only IS-A relations are found (at least in our dictionaries).

<sup>4</sup> In our dictionaries there is no explicit statement about the difference among reflexive verbs, inherently reflexive verbs and those ergative verbs which have a reflexive form (cf. Burzio (1986: 36 ff.)). Furthermore, the terminology used in the two Italian sources is different; however, for the time being, we have maintained the terminology which we find.

<sup>5</sup> This information is extracted by analysing the definition(s) given for the genus term itself in the dictionary being used. Then, after building taxonomies automatically, for each genus term we have to choose manually (there is no possibility to do it automatically as far as our mono-lingual Italian dictionaries are concerned) in which sense, among those found within its definitions, the genus term is used in the definitions of its hyponyms.

<sup>6</sup> Actually, a PP with the preposition *a* may have other semantic values; however, within motion verb taxonomies extracted from our dictionaries it may only have the two values shown.

<sup>7</sup> Talmy (1985: 141) distinguishes between "translational" and "contained" motion: "in the former, an object's basic location shifts from one point to another in space. In the latter, the object keeps its same basic, or 'average' location". As a matter of fact, when he says that Romance languages cannot conflate Motion and Manner, he rather seems to refer to "translational motion".

<sup>8</sup> It is necessary to say that within DMI definition of *passare* there is no reference to a PURPOSE. However, we are using different mono-lingual sources in order to overcome the limits which each single source present and that have often been emphasized; so when we find information in one dictionary which is missing in the other, we generally keep it and consider it as more significant than its lack.

## References

Alonge, A. (1991), "Extraction of Information on Aktionsart from Verb Definitions in Machine-Readable Dictionaries", in *Proceedings of the*

*11th International Workshop on Expert Systems & their Applications*, Avignon.

Boguraev, B. & Pustejovsky, J. (1990), "Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design", in *Proceedings of the 13th International COLING*, Helsinki.

Burzio, L. (1986), *Italian Syntax*, Dordrecht, Reidel.

Calzolari, N. (1984), "Detecting patterns in a Lexical Database", in *Proceedings of the 10th International COLING*, Stanford (Calif.).

id. (1991), "Acquiring and Representing Semantic Information in a Lexical Knowledge Base", in Pustejovsky, (ed.) *Lexical Semantics and Knowledge Representation. Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the ACL*, Berkeley, Cal.

Chodorow, M. S., R. J. Byrd and G. E. Heidorn (1985), "Extracting semantic hierarchies from a large on-line dictionary", in *Proceedings of the 23rd ACL Annual Conference*, University of Chicago.

Dowty, D. R. (1988), "Thematic Protoroles, Subject Selection, and Lexical Semantic Defaults", 1987 LSA Colloquium Paper

Levin B. (1985) (ed.), *Lexical Semantics in Review*, Lexicon Project WPs, 1, Cambridge, MIT.

Levin, B. and M. Rappaport (1991), "The Lexical Semantics of Verbs of Motion: the Perspective from Unaccusativity", ms.

Perlmutter, D. M. (1978), "Impersonal Passives and the Unaccusative Hypothesis", BLS, 4.

Pustejovsky, J. (1989), "Current Issues in Computational Lexical Semantics", in *Proceedings of the 4th Conference of the European Chapter of the ACL*, Manchester, England.

Sanfilippo, A. (1991), "LKB Encoding of Lexical Knowledge from Machine-Readable Dictionaries", ms.

Talmy, L. (1985), "Lexicalization Patterns: Semantic Structure in Lexical Forms", in T. Shopen (ed.) *Language Typology and Syntactic description* 3, Cambridge University Press.

Vendler, Z. (1967), *Linguistics in Philosophy*, Ithaca, Cornell University Press.