Antonietta Alonge, Università di Pisa, Dip. di Linguistica

Machine-readable dictionaries and lexical information on verbs

ABSTRACT: This paper reports some of the results of a research which is being carried out within the ESPRIT project ACQUILEX, aiming at demonstrating the feasibility of building a multilingual lexical knowledge base (for NLP systems) by exploiting different mono- and bi-lingual machine-readable dictionaries as sources of lexical data. In particular, this paper is concerned with the work being done in order to develop techniques and methodologies for the semi-automatic extraction of semantic and syntactic information on Italian verbs from dictionary definitions.

1. Introduction

In recent years, since the need for large computational lexicons has become a major concern for researchers working in the field of NLP, many (computational) lexicographers / linguists have turned to machine-readable dictionaries (MRDs) as potentially reusable sources of lexical data (cf. Boguraev & Briscoe 1989). Within the ESPRIT project ACQUILEX[1] the possibility of developing a multilingual and maximally reusable (by different researchers, for different purposes) lexical knowledge base (LKB) for NPL systems, utilising existing MRDs of four languages as sources of data, is being explored and our first results seem to be positive. As a matter of fact, not all the necessary lexical information is contained in dictionaries and, furthermore, there are big differences among dictionaries themselves. In particular, as fas as semantic and syntactic information is concerned, within the DIZIONARIO DELLA LINGUA ITALIANA Garzanti (GRZ) and the DIZIONARIO-MACCHINA DELL'ITALIANO (DMI – a MRD mainly based on the Zingarelli VOCABOLARIO DELLA LINGUA ITALIANA), which are the monolingual dictionaries used for the project in Pisa[2], we mainly find semantic information (and this is only implicitly available), while the LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH, which is being used both in Cambridge and Amsterdam, contains additional syntactic information and also explicitly coded semantic information. Notwithstanding, we think that some very important lexical information can be semi-automatically extracted from our MRDs with a significant saving both in resources and in time compared to coding the same information by hand.

Althoug lexical knowledge encompasses phonological, morphological, syntactic and semantic knowledge, within our research we have been interested mainly in the extraction of semantic and syntactic infomation on verbs from MRDs. In this paper, therefore, the work being carried out in order to identify and extract these kinds of data will be described; in particular, we shall report on the methodologies we developed and are developing in order to extract not only the information which is contained in the first part of verb definitions (where superordinate categories – "genus terms" – of the entry are generally found), but also data contained within the "differentia" part (where what distinguishes an instance of a genus term from other instances is stated) which had not really been exploited before (but cf. Calzolari 1984; 1991 for proposals related to the analysis of this part of definitions). 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 (which is recognized by all the most recent theories on natural language) and that NLP systems need lexical information. 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; 1989; Levin & Rappaport 1991). Finally, we tried to follow some indications, provided in works such as those of Pustejovsky (1989) of 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 various 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; qualia structure (see Pustejovsky, 1989); lexical inheritance structure.

2. Dictionary definitions and verbs

As Amsler (1980) first showed by manually extracting and disambiguating genus terms from a pocket dictionary, it is possible to build taxonomic (or IS-A) hierarchies by analysing dictionary definitions. Procedures aimed at building IS-A hierarchies from dictionary definitions semi-automatically are now well established (Calzolari 1984; 1988; Chodorow et al. 1985), and work has also been done on other kinds of relations found in the first part of a definition (Vossen et al. 1989; Alonge et al. 1991). However, besides utilizing procedures for building taxonomies and exploiting the information which they provide, we are also carefully analysing the data which are found in the differentia part of definitions, by means of pattern-matching procedures to be applied to the output of the syntactic analysis of definitions.

Verb dictionary definitions are perhaps not as rich as noun definitions; only IS-A relations are found and the differentia often consists of an adverbial phrase or little more. In any case, by analyzing taxonomies, we have already been able to extract some important information, especially using the possibility of having inheritance of information as a consequence of the IS-A link (on this issue, cf. Copestake 1990); moreover, some important data have been found in the differentia, too.

3. Information extracted on verbs

As already mentioned, in our dictionaries we mainly find semantic information; that is, as far as syntactic information is concerned, we may only know if a verb is transitive, intransitive or reflexive[3]; when it is intransitive, we may determine if it is unaccusative or unergative by taking into consideration the auxiliary selected (indicated within Collins dictionary) because Italian unaccusative verbs take *essere* (to be)[4]. However, the semantic information which we are able to extract can help us identify also syntactic characteristics of the verbs being analysed, as will be shown below.

3.1. Aktionsart

The enormous number of works which have been devoted to Aktionsart (or "lexical aspect") by linguists is obviously due to the importance of this notion in the description of verb semantics, but also syntax. The classification of verbs according to Aktionsart, in fact, has important syntactic consequences in that it determines the possibility of a given verb occuring with specific adverbial phrases, verbs, etc. (Vendler 1967; Dowty 1979) and it also seems to determine the syntactic realization of arguments (Tenny 1988). Furthermore, the Aktionsart-class to which a verb belongs has consequences at the level of discourse (Dowty 1986). Therefore, a research was carried out in order to semi-automatically classify verbs utilizing MRDs (cf. Alonge 1991). After classifying some genus terms according to Vendler's (1967) proposal (distinguishing among *states, activities, accomplishments* and *achievements*), we tried to see if the hyponyms of a genus term verb shared with it the Aktionsart-class; i.e., we looked for evidence of inheritance of Aktionsart-classification along IS-A hierarchies. This was indeed the case for about 90% of the entries considered (nearly) 4,000); however, we could classify almost all the other verbs considered by taking into account also the differentia of definitions (by means of a pattern-matching procedure) because Aktionsart-classification actually pertains to whole VPs, and internal arguments of verbs may determine different classifications of VPs containing the same verb. To give just two examples, *scorrazzare* (to rove about) is defined, in GRZ (sense) 1, as "correre qua e là" (to run about); since *correre* is an activity verb and what we find within the differentia cannot yield a different classification of the VP, then *scorrazzare* was classified as an activity verb, too. On the other hand, *accorrere* (to rush to someone) is defined, GRZ 1, as "correre verso qualcuno" (to run towards someone) and, although *correre* is an activity, *accorrere* had to be classified as an achievement, because in its definition the genus term occurs with a PP which makes the whole VP behave as an accomplishment or an achievement (cf. Dowty, 1979).

3.2. Components of meaning, typical subjects and thematic roles

With the main goal of overcoming the well-known limits that individual dictionaries present (incoherences, lack of data, etc.), we decided to merge the information coming from our two sources. By analysing some taxonomies, we identified groups of them which could be associated under a same "conceptual label". The main reasons for doing this were that we had found many groups of genus terms which were circularly defined in both sources and, morever, words which were found to be hyponyms of one genus term of a group in one dictionary could be found in the taxonomy of another genus term (of the same group) in the other. The following are some of the taxonomies which were associated:

- MOVE : Muoversi (intransitive *to move*); Muovere (both transitive and intransitive *to move*); Andare (*to go*)
- MAKE : Rendere (*to make*); Far-Diventare / Divenire (*to cause to become*)
- BECOME : Diventare / Divenire (*to become*)
- CAUSE : Causare; Provocare; Cagionare; Procurare; Arrecare; Produrre (they all can be translated as *to cause*)

As a matter of fact, the "conceptual labesl" which were used to associate taxonomies turn out to correspond to "semantic primitives" indicated in various linguistic theories and they can be used to state the basic component of meaning which large subsets of words share and which may determine important syntactic behaviours for verbs in the same subset. As Levin (1985, 1989) showed, there is, in fact, a direct correspondence between the syntactic behaviour of groups of verbs and elements of meaning which they have in common: e.g., diatheses alternations (i.e., alternations in the expression of the arguments of verbs), which classes of verbs display, seem generally connected with specific components of meaning. Therefore, e.g., by grouping together some taxonomies under the labels MAKE and BECOME, it was possible to identify verbs displaying the so-called "causative-inchoative alternation" (with the causative variant of the verb occurring in the MAKE taxonomy, and the inchoative variant under BECOME): these are verbs which may be used as causative-transitive verbs or inchoative-intransitive ones, as may be seen for *imbiancare* (to whiten) in (1) and (2) below:

1. Il tempo ha imbiancato i capelli di Maria. [Time has whitened Maria's hair.)
2. I capelli di Maria sono imbiancati. [Maria's hair has whitened.)

(For details on a research carried out on DMI in relation to these taxonomies, see Antelmi & Roventini 1991.) Similar connections can be drawn among other basic components of meaning shared by verbs found within the same taxonomy and the syntactic behaviour of the verbs themselves.

By analysing the differentia part of definitions we then extracted some other important information again connected with components of meaning; moreover, we often found indications on typical subjects of the verbs defined. This was done by identifying recurrent patterns, clearly referring to specific semantic categories, within definitions of verbs occurring in the same taxonomies (or in taxonomies which had been "associated"). I.e., we first examined manually some of the definitions of verbs within the same taxonomy and individuated 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, obviously, found in each definition. The following are examples of the patterns found within the definitions in the taxonomies of *colpire* and *muoversi* and of the components of meaning which were connected with them:

COLPIRE (to hit):

- WITH_INSTR: *con NP*
- GOAL: NP (direct object of the genus term)
- MANNER: AdvP
- ITERATION: AdvP
- PURPOSE: *per VP*
- TYPICAL SUBJECT: *detto di / si dice di / di NP*

MUOVERSI:

- MANNER OF MOVEMENT: *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 of definitions and relates them to semantic categories (or typical subjects), by taking into consideration the taxonomy in which the entry defined is found. In fact, similar patterns may indicate different semantic components in relation to different taxonomies. E.g., the pattern *con NP* (with NP) indicates an INSTRUMENT in relation to the taxonomy of *colpire*, but it indicates a MANNER OF MOVEMENT when found within definitions of motion verbs[5]. Therefore, it is necessary to develop different procedures for each taxonomy (or groups of taxonomies associated under a same label). Furthermore, sometimes different semantic categories related to the same taxonomy may be indicated by the same lexical category / pattern, so that it becomes 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[6]. The components of meaning individuated by means of this analysis may be used to derive useful information on (classes of) verbs. For instance, the data extracted on movement verbs were used to further classify these verbs according to a proposal by Levin & Rappaport (1991). Even if we speak of *one* class of motion verbs, the authors emphasized that these verbs do not constitute a linguistically significant natural class: actually, as far as *intransitive* verbs of motion are concerned, three classes of verbs can be identified: 1) *arrive* verbs; 2) *run* verbs; 3) *roll* verbs.

Each class seems to be characterized by a particular component of meaning, which also determines the status of a verb as unaccusative or unergative (on the "Unaccusative Hypothesis" cf. Perlmutter 1978 and Burzio 1986):

1) DIRECTION (GOAL) → unaccusative;
2) MANNER + PROTAGONIST CONTROL (i.e., the moving "object" causes the movement) → unergative;
3) MANNER + NO PROTAGONIST CONTROL (i.e., there is a direct external cause for the movement) → unaccusative.

By taking into consideration the differentia of definitions (where we find information about components of meaning such as GOAL and MANNER, but not about the presence of a control on the part of the protagonist of motion) as well as information on the auxiliary selected by a verb, which allows us to classify a verb as unaccusative or unergative, we were able to classify motion verbs even further. 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 project (cf. Sanfilippo 1991). Dowty individuates 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.

a) The prototypical agent of a verb is the thematic relation associated with the argument having the highest number of proto-agent properties entailed by the meaning of the verb and inherited by default
b) The prototypical patient is the tematic relation associated with 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) (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 can be identified through the analysis of definitions.

By examining the verb 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 which corresponds to the "moving object" and for which either the manner of movement 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 actually inherited from the genus term *muoversi*; the specification relative to the manner of movement 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 movement 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)

Therefore, within the LKB which is being developed, *andare* will be described as related to an argument bearing a proto-agent-move-path role, while *oscillare* will be connected with a proto-agent-move-manner.

4. Conclusion

By combining theoretical assumptions with empirical observations we have been able to extract some important semantic, and also syntactic, information on verbs. The work is still in progress and we are automating the various stages of the analyses; a similar work will then be carried out in relation to other taxonomies extracted from our MRDs.

Endnotes

[1] ESPRIT BRA-3030, on the "Acquisition of Lexical Knowledge for Natural Language Processing Systems". Universities of Amsterdam, Barcelona, Cambridge, Dublin and Pisa research teams are collaborating within this project.

[2] In Pisa, we also utilize the Collins bilingual (English-Italian) dictionary.

[3] Actually, Burzio (1986) distinguishes among reflexive verbs, inherently reflexive verbs and ergative verbs which have a reflexive form. In our dictionaries two groups of reflexive verbs are identified (and different terminologies are used in the two sources) and, for the time being, we have extracted the information which is found, without further analysis.

[4] As a matter of fact, also reflexive forms of ergative verbs and inherently reflexive verbs are unaccusative. However, since we do not find clear data on these classes within dictionaries (cf. above), we classified manually unaccusative reflexive verbs.

[5] This pattern could also refer to other semantic categories, but generally only relevant information is given in dictionary definitions and we saw that, within these taxonomies, this PP is used only to refer to the components of meaning indicated above.

[6] In any case, such lists are restricted because of the characteristics of dictionaries; in fact, as Calzolari (1991, 189) points out: "the lexicographic tradition has exerted a (usually unconscious) control over the defining vocabulary (...) and the schemata of defining formulas".

Bibliography

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.

ALONGE, A., CALZOLARI, N., HAGMAN, J., MARINAI, E., MONTEMAGNI, S., PETERS, C., PICCHI, E., ROVENTINI, A., SPANU, A., & ZAMPOLLI, A. (1991): "An Overview of Work on Semantic Taxonomies in Pisa". In: Semantic Taxonomies, Esprit BRA-3030 ACQUILEX Deliverable # 2.3.8.

AMSLER, R. (1980): The Structure of the Merriam-Webster Pocket Dictionary, Doctoral Dissertation, University of Texas, Austin.

ANTELMI, D. & ROVENTINI, A. (1991): "Semantic Relationships within a Set of Verbal Entries in the Italian Lexical Database". In: Proceedings of the 4th Euralex Congress, Benalmádena, Malaga.

BOGURAEV, B. & BRISCOE, T. (1989): Computational Lexicography for Natural Language Processing, Harlow Longman.

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.).

CALZOLARI, N. (1988): "The Dictionary and the Thesaurus can be combined". In: Relational Models of the Lexicon. Ed. by M. W. Evens. Cambridge (Mass.), CUP.

CALZOLARI, N. (1991): "Acquiring and Representing Semantic Information in a Lexical Knowledge Base". In: Proceedings of the Workshop on Lexical Semantics. Ed. by J. Pustejovsky. Berkeley.

CHODOROW, M. S., BYRD R. J. & HEIDORN G.E. (1985): "Extracting Semantic Hierarchies from a Large On-Line Dictionary". In: Proceedings of the 23rd ACL Annual Conference, Chicago.

COPESTAKE, A. (1990): "An Approach to Building the Hierarchical Element of a Lexical Knowledge Base from a Machine Readable Dictionary", Esprit BRA-3030 ACQUILEX WP No. 8.

DOWTY, D. R. (1979): Word Meaning and Montague Grammar: the Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ, Dordrecht, D. Reidel.

DOWTY, D. R. (1986): "The Effects of Aspectual Classes on the Temporal Structure of Discourse: Semantics or Pragmatics?". In: Linguistics and Philosophy, 9,1.

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

LEVIN, B. (ed.) (1985): Lexical Semantics in Review. Lexicon Project WPs, 1, Cambridge (Mass.) The MIT Press.

LEVIN, B. (1989): English Verbal Diathesis. Lexicon Project WPs, 32, Cambridge (Mass.), The MIT Press.

LEVIN, B. & 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.

TENNY, C. (1988): "The Aspectual Interface Hypothesis: the Connection between Syntax and Lexical Semantics". In: Studies in Generative Approaches to Aspect. Ed. by C. Tenny, Lexicon Project WPs 24, Cambridge, Mass., MIT.

VENDLER, Z. (1967): Linguistics in Philosophy, Ithaca, Cornell University Press.

VOSSEN, P., MEIJS W. & den BROEDER M. (1989): "Meaning and Structure in Dictionary Definitions". In: Boguraev, B. and T. Briscoe (1989).

KEYWORDS: Machine-readable Dictionaries, Lexical Databases, Lexical Knowledge Base, genus term, differentia.