Towards a Sensorimotor WordNet^{SM}: Closing the Semantic Gap*

Gutemberg Guerra-Filho and **Yiannis Aloimonos**
Computer Vision Laboratory
Department of Computer Science
University of Maryland
College Park, MD 20742
guerra@cs.umd.edu yiannis@cfar.umd.edu

Abstract

We have empirically discovered that the space of human actions has a grammatical structure. This is a motoric space consisting of the evolution of the joint angles of the human body in movement. Furthermore, the process of assembling individual human movements into higher level descriptions resembles in a natural sense the process of speech recognition. Thus the space of human activity has its own phonemes, morphemes, words (verbs, nouns, adjectives, adverbs), and sentences formed by its own syntax. This has a number of implications for the grounding problem and cognition in general. With regard to WordNet, the theory points to a future Sensorimotor WordNet which contains a map between the nodes of the current WordNet and the space consisting of human action. In this paper, we suggest initial steps towards closing the semantic gap by grounding language with visuomotor information. The grounding takes place on a set of primitive words which are selected here through verb classification of the WordNet lexicon. A formal approach to the identification of primitive words would consider the basic atoms of WordNet extensions. However, one further extension is required to incorporate grounded information into WordNet in the direction of a sensorimotor WordNet, designated here as WordNet^{SM}.

Introduction

Computational models of Natural Language Processing are essentially symbolic systems considering lexical and semantic information. WordNet is a general-purpose lexical database which organizes a vocabulary into synonym sets and a semantic network, Fellbaum (1998), Miller *at al.* (1990). In WordNet, synsets are words in the same lexical category expressing the same meaning. The definition of a synset is a conceptual gloss described in natural language. Extensions of WordNet intend to use logical predicates and axiomatic theory to formally specify synset definitions, Gangemi *at al.* (2003), Harabagiu *at al.* (1999). These extensions increase the reasoning and manipulative power of WordNet which is still a pure symbolic system.

The semantic interpretation of a symbolic representation system, such as natural language, cannot be based only on meaningless arbitrary symbols. The symbol grounding problem, Harnard (1990), addresses this semantic gap and suggests that the primitives of a formal symbolic system should be associated with grounded representation connected to physical experience in the world. Once a grounded set of elementary symbols is provided, higher-order symbols of a language are generated by composition and the semantic gap is closed by the intrinsic grounding of the elementary set.

The existence of mirror neurons in humans suggests the use of the same representation for perceptual and motor tasks. Mirror neurons would activate when a subject performs a specific action. The same neurons will also fire when the subject observes the same action.

A grounded representation is a sensorimotor projection of objects and events to which elementary symbols refer. In this paper, we concentrate in events associated with human activities. This way, a sensorimotor projection consists in the translation from a non-symbolic analog representation of human activities in the world to a grounded non-arbitrary symbolic representation according to invariant features which allow cognitive tasks such as recognition. The sensorimotor projection of primitive words leads to language grounding. Language grounding for verbs has been addressed by Siskind (2001) and Bailey *at al.* (1997) from the perspective of perception and action, respectively. Roy (2005) introduces a theoretical framework for grounding language using semiotics and schema theory.

A sensorimotor representation for the primitive words grounds the meaning in perception and action. Such representation allows modeling of complex multi-modal phenomena and the understanding of situated language acquisition with applications to conversational machines.

In this paper, we apply verb classification to the selection of an elementary set of human actions from WordNet verbal data. This is one possible approach to find a primitive set of verbs, denominated concrete verbs, for language grounding. Another possibility would consider the basic atoms of the formal framework of WordNet extensions.

We designed and implemented a new semi-automatic verb classification system based on WordNet. The system was used to identify a human activity lexicon. The classification process uses the verb hierarchy constructed from the troponymy relation. A domain specific verb classification process was performed in the WordNet verbal data in order

* The support of ARDA (under the VACE program) and of NSF (under the HSD program) are gratefully acknowledged.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 171–176.

to select concrete verbs for visuo-motor grounding. We selected 1471 synsets organized in a hierarchical forest of 23 trees.

We propose the extension of WordNet towards the aggregation of sensorimotor data connected to the set of concrete verbs. This extension, denominated here as WordNet^{SM}, is suggested for grounding language with a **S**ensori**M**otor representation. In this paper, we present a visuo-motor language as the grounded representation for the concrete verbs which constitute the language primitives. Our visuo-motor language is called Human Activity Language (HAL). HAL is specified in a linguistic approach, where phonetics, morphology and, syntax are defined, Guerra-Filho *at al.* (2005). The linguistics framework is used to represent human movement with a symbolic, but non-arbitrary, system. A linguistic approach benefits from the theory of Automatic Speech Recognition and Natural Language Processing.

Section 1 discusses related work to verb classification. In Section 2, we present our semi-automatic verb classification system. Section 3 describes HAL and the initial steps towards WordNet^{SM}: a sensorimotor WordNet.

1 Related Work

The foundations for verb classification have been established on semantic verb classes for English, Levin (1993), and Spanish, Vázquez *at al*. (2000). Verb classification provides lexical organization which can capture generalization and specialization over verbs. Automatic verb classification is based on the linguistic hypothesis that semantic properties of verbs are reflected in their syntactic behavior, Levin (1993). This hypothesis is valid only to a certain extent when considering the choice of verb arguments. On the other hand, manual classification of verbs is a difficult and resource intensive task, Miller *at al*. (1990). Here, we suggest a semi-automatic approach to perform verb classification according to a specific domain: observable movement.

Corpus-based approaches to automatic verb classification use statistical features over the syntax of verbs as training data for an automatic classifier. Merlo and Stevenson (2001) report on supervised learning experiments to automatically classify English verbs into three optionally intransitive verb classes (unergative, unaccusative, and object-drop verbs) based on their predicate-argument structure and statistics from large annotated corpora about five features. This statistical corpus-based method uses theoretical linguistic properties of the thematic roles assigned by the verbs. Schulte im Walde and Brew (2002) obtain German verb classes automatically using a robust statistical parser (k-means clustering) based on syntactic descriptors. In these methods, a discriminating set of features is determined manually. Joanis and Stevenson (2003) develop a general feature space for automatic verb classification. The general feature space avoids the need for individual manual development of features for specific classes. In the corpus-based approaches, the features extracted from a corpus are noisy and only indirect indicators of semantics, while relevant semantic properties are not expressed openly.

In WordNet, verb meaning is represented in terms of semantic relations rather than semantic primitives, Miller and Fellbaum (1991). WordNet has 15 semantic domains for verbs, but none of those domains represent observable movement closely. Information relevant for the lexical encoding of verbs is domain specific and is missing in a general-purpose classification like WordNet. This way, in the linguistic literature there is no verb classification that captures the observable movement domain from a visual recognition perspective.

2 Verb Classification System

We define a human action as consisting of visually observable movements. This way, a human action lexicon is restricted to the visually perceived motor domain.

The action lexicon may be classified according to the environment where a human acts. The environment is basically the *stage* where a person interacts, and intrinsically defines all possible actions that may be performed, similarly to the theory of affordances, Gibson (1979). The stage concept gives rise to a classification methodology of human actions according to the objects and actors present in the stage.

The stage may be set up initially with a single actor and the ground. The *ground class* (GND) includes human actions that require only the ground to be performed by the actor. Examples of verbs in the ground class are walk, jump, nod, and clap.

A second class of actions involves a general object. A *general object* is any object that has no specific feature or property required by the action. This way, the actions found in this class are related to manipulation activities. Actions in the *general object class* (GOB) are touch, push, strike, and prehend, among others.

Another class involves the actor's interaction with a general person. A *general person* is any person with no specific status, position, or skill. The *general person class* (GPE) includes activities involving two people interacting with each other. Verbs in this class are, for example, caress, pursue, pass, and embrace.

A *specific object* is an object with particular properties and functionality. This functionality is used when specific actions are performed. The *specific object class* (SOB) includes verbs such as comb, brush, write, and stamp.

Similarly, a *specific person* is a person who performs a particular function which requires status, position, or skill. The *specific person class* (SPE) includes verbs representing actions where the actor is a specific person or interacts with a specific person who plays a particular role while performing some specialized function. Examples of verbs in this class are confess (with a priest), arrest (by a policemen), sentence (by a judge), and diagnose (by a doctor).

A last class includes verbs corresponding to actions performed by a group of people. This way, there is no single actor executing the activity but many actors which form a group. Verbs in the *group class* (GRP) are, for example, riot, play, line up, and assemble.

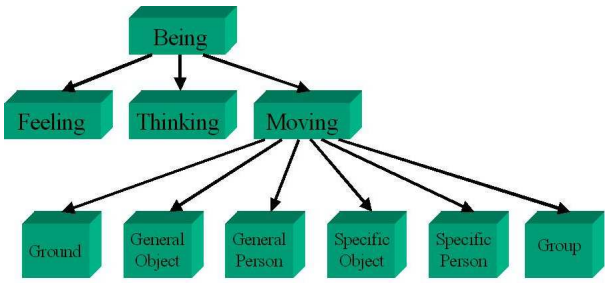

Figure 1: Motor domain verb hierarchy.

In WordNet, a network links words according to semantic relations. The *troponymy* relation is a special case of entailment where the troponym and the more general verb are always temporally coextensive. A verb hierarchy organized as a forest of trees of synsets is constructed from the troponymy relation.

The classification process takes advantage of this hierarchical organization of verb synsets. The classification starts at the highest level and continues towards the lower levels. At each level, the synsets are manually categorized as any of the classes in the stage concept; otherwise an unknown tag is assigned to the synset. After a synset is manually categorized, its descendents are automatically classified within the same class. This semi-automatic method greatly reduced the amount of manual work required to classify the whole verb vocabulary, since most of the verbs are automatically classified during the descendents visit.

Once the verbs are classified, a consistency check is performed using the entailment and antonymy relations. We assume that two synsets related by entailment or antonymy should be in the same class.

After the semi-automatic initial classification phase, the synsets tagged as unknown verbs are categorized according to the classification of its descendents. This automatic phase starts at the unknown verbs with all children already classified as one of the six stage classes. A final verification of the classification is performed on the synset leaves of the verb hierarchy. At each leaf synset, the class is manually checked.

The concrete verbs resulting from this process represents a lexicon that is organized hierarchically and divided into stage classes. The ground class, general object class, and general person class contain 1471 synsets organized in a hierarchical forest with 23 trees. These concrete verbs represent an initial effort to identify the elementary set of verbs for language grounding[1].

3 Human Activity Language

A language consists of a subsystem that selects certain states among all possible states (phonology), a subsystem for making words (morphology), and a subsystem for making sentences out of words (syntax).

[1]The complete list of verbs obtained is found at `http://www.cs.umd.edu/~guerra/concreteVerbs.html`

We captured videos featuring 90 different human activities and the corresponding three-dimensional reconstruction for trajectories of body parts was found using our own Motion Capture system, Guerra-Filho (2005). Given this three-dimensional reconstruction, joint angles were computed to describe human movement. The joint angles are the original 3D representation corresponding to the analog signal which is translated into our grounded representation: the Human Activity Language.

3.1 Phonology

In a visuo-motor language, a phonetic description consists of a sequence of articulatory configurations. The HAL phonology is based on invariant features of first derivatives (velocity) and second derivatives (acceleration) of joint angles (see Fig. 2).

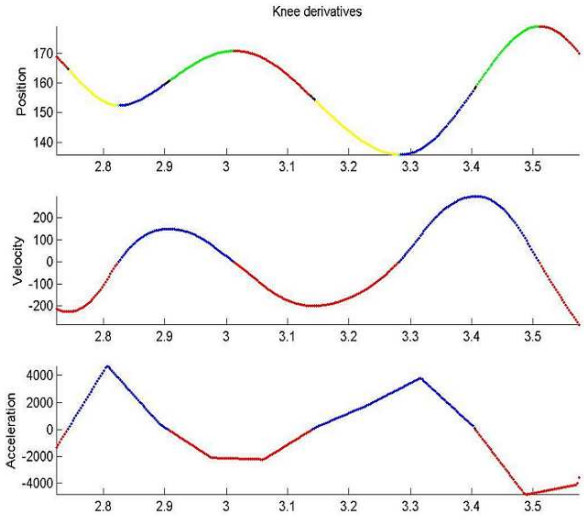

Figure 2: Knee angle derivatives during jog.

The joint movement is segmented according to the sign of the angular derivatives. We use six possible atomic states: four combinations of positive and negative signs for the two derivatives (R, Y, B, G), one state for zero acceleration (V), and another state for zero velocity. Each segment has an initial joint angle φ and a final angle θ. The *potential* Δ of a segment is the quatized absolute difference between these two angles.

An alphabet of atomic joint movements is necessary for a grounded symbolic representation of action. Each segment corresponds to an atom $\alpha\Delta$, where α is a symbol associated with the segment's state and Δ corresponds to the segment's potential. This way, the string R1 Y3 B2 G3 R4 Y6 B7 G6 R2 corresponds to the activity jog according to the knee joint (see Fig. 3).

3.2 Morphology

Given the phonological representation for an activity lexicon, a hierarchical organization is derived in the form of morphological grammars for the activity lexicon. The

jog := B0 G0 R2 Y3 B3 V0 G3 R0 Y0
jump := G0 R0 Y0 B2 G1 V0 B5 G3 R0 Y0 R10 Y5 B0 G0 R0 Y0 B0 G1 B0 G0 R0 Y1 B0 V0 B0 V0 G0 R0 Y0 B0 G0 R0 Y0 V0
run := B4 G7 R1 Y2 B1 G0 R3 Y5 V0 Y0
scuff := **R0 V0 Y0 V1 Y0 V0 R0 V0 Y0** V0 R0 V0 Y0 B0 V0 G0 V1 G0 V0 G0 □0 V0 □0
stomp := G0 R0 Y0 V0 R1 V0 Y2 V1 R0 V0 Y1 V0 Y0 B0 G0 V0 B1 V0 G1 V0 B1 V0 G2 R2 Y0 B0
swagger := Y0 V0 R1 V0 Y0 V0 R0 V0 Y0 V0 □0 R0 V0 Y0 V0 R0 V0 Y0 B2 G2 V0 B0 V0 G0 V0 B0 V0 G0 V0 G0 R1
tiptoe := B0 G1 V1 B0 V0 G1 □0 B0 V0 G0 **R0 V0 Y0 V0 Y0 V0 R0 V0 Y0** B0 V0 G0 R0 V0 Y0
toe := R0 V1 R0 V0 Y0 V0 R0 Y0 B0 G1 V2 G1 R0 V0 Y0 B0 G0
troop := R0 Y0 B0 G0 R1 V3 Y3 V1 Y0 B5 G3 R0 Y0 B2 G3 R2 V0 Y2 B0 G0
walk := **R0 V0 Y0 V1 Y0 V0 R0 V0 Y0** B2 G2 V0 B0 V0 G0 R0 V0 Y0 □0 R0 Y1 V0 R0 V0 Y1

Figure 4: Lowest-level morphological grammar.

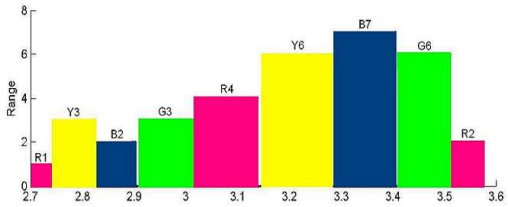

Figure 3: Segmentation of the knee joint angle.

phonological strings for each activity represent the lowest level in the morphological grammar (see Fig. 4).

The generation of a higher-level morphological grammar involves finding common substrings in different activities of the lexicon. Our algorithm finds the most frequent pair $\alpha_i \Delta_i \ \alpha_j \Delta_j$ of consecutive atoms in the current grammar. A new grammar rule $L_n := \alpha_i \Delta_i \ \alpha_j \Delta_j$ is then created. A higher-level grammar is generated using the new rule. Each occurrence of the pair of atoms $\alpha_i \Delta_i \ \alpha_j \Delta_j$ in the current grammar is replaced by a non-terminal L_n. This process is repeated until the most frequent pair in the current grammar has less than two occurrences and, consequently, the highest level of the grammar is reached. The highest level of the grammar contains the lexical units (words) of HAL.

3.3 Syntax

The Subject-Verb-Object (SVO) pattern of syntax is a reflection of cause and effect: something doing something to something else. In most languages, the sequence of signals falls into a subject/predicate pattern. An action is represented by a word that has the structure of a sentence: the agent or subject is a set of active body parts; the action or verb is the motion of those parts. In many such words, the action is transitive and involves an object or another patient body part.

In a sentence, a noun represents the subjects performing an activity or objects receiving an activity. A noun in a HAL sentence corresponds to the body parts active during the execution of a human activity and to the possible objects involved passively in the action.

The initial posture for a HAL sentence is analogous to an adjective which further describes (modifies) the active joints (noun) in the sentence. The HAL adjective is represented by a string of integers considering only the active joints in the activity. Each element in this string corresponds to the initial angle of an active joint.

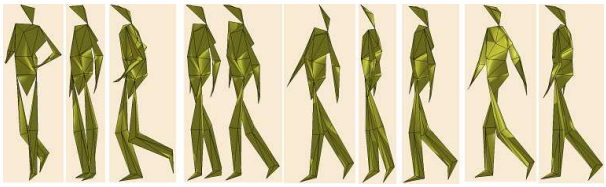

Figure 5: HAL adjectives: initial postures.

The sentence verb represents the changes each active joint experiences during the action execution. The representation for a HAL verb was discussed previously. However, further description is required to deal with coordination among different joints.

A coordinated segment is a time interval delimited by events representing local minima and maxima in the joint angle function for any of the active joints. These events occur in between specific atomic pairs ($\mathrm{Y}\Delta$ $\mathrm{B}\Delta$ and $\mathrm{G}\Delta$ $\mathrm{R}\Delta$) and, consequently, may be computed from the HAL verb strings (see Fig. 6).

A HAL adverb is a string of multiplicative constants modeling the variation in the execution time of each coordinated segment. A HAL adverb is appended to a verb in such a way that each value in the adverb string corresponds to a coordinated segment in the verb.

A HAL sentence S := NP VP consists of a noun phrase (noun + adjective) and a verbal phrase (verb + adverb), where NP := N Adj and VP := V Adv. The organization of human movement is simultaneous and sequential. This way, the basic HAL syntax expands to two orthogonal axes based on joint (parallel syntax) and time (sequential syntax) structure (see Fig. 7).

The parallel syntax concerns simultaneous activities represented by parallel sentences $\mathrm{S}_{t,j}$ and $\mathrm{S}_{t,j+1}$ and constrains the respective nouns to be different: $\mathrm{N}_{t,j} \neq \mathrm{N}_{t,j+1}$. This constraint states that simultaneous movement must be performed by different body parts.

The temporal sequential combination of action sentences ($\mathrm{S}_{t,j}$ $\mathrm{S}_{t+1,j}$) must obey the cause and effect rule. The HAL noun phrase must experience the verb cause and the joint configuration effect must lead to a posture corresponding

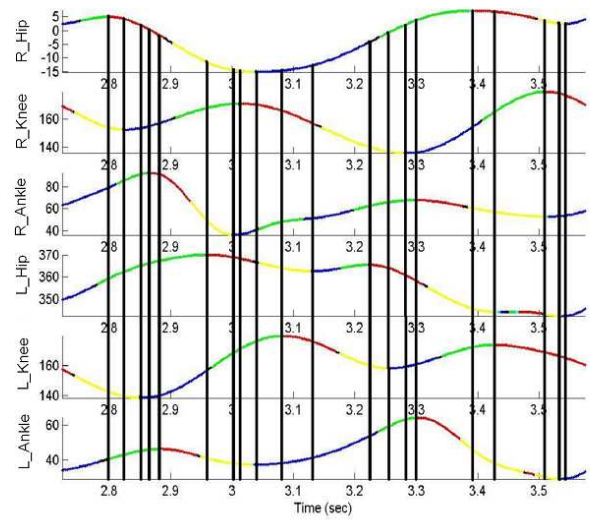

Figure 6: Coordinated segments.

to the noun phrase of the next sentence. Considering noun phrases as points and verb phrases as vectors in the same space, the cause and effect rule becomes $\mathrm{NP}_{t,j} + \mathrm{VP}_{t,j} = \mathrm{NP}_{t+1,j}$. The cause and effect rule is physically consistent and embeds the ordering concept of syntax.

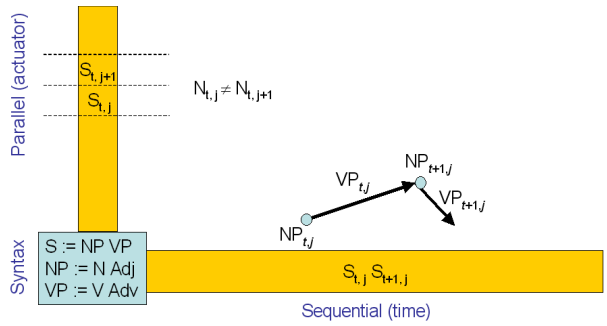

Figure 7: HAL syntax.

Conclusion

According to a motor domain verb hierarchy, the stage concept categorizes human actions based on the objects and actors in the environment. Six classes were suggested to categorize verbs in the motor domain. We designed and implemented a semi-automatic verb classification system based on WordNet. The system was used to identify a human activity lexicon with 1471 synsets organized in a hierarchical forest of 23 trees.

A visuo-motor language was presented using a linguistic approach by specifying phonology, morphology, and syntax of the visuo-motor information of human actions. In phonology, we introduced basic atomic segments that are used to compose human activity. Segments are characterized according to the sign of the first and second angular derivatives of joints. In morphology, we studied how HAL phonemes were combined to form strings representing human activity. Basically, we explored common substrings to generate a higher-level morphological grammar which is more compact and suggests the existence of lexical units working as visuo-motor subprograms.

In syntax, we presented a model for visuo-motor sentence construction where the subject in a sentence corresponds to the active joints (noun) modified by a posture (adjective). HAL verbs represent the changes each active joint experiences during an activity execution. Coordination among different joints is specified by timing the atomic segments and appending elastic discrete values (adverb) to coordinated segments. The adverb is used to adjust and modify the action execution.

In this paper, we suggested initial steps towards closing the semantic gap by grounding language with sensorimotor information. The grounding takes place on a set of primitive words which were selected here through verb classification of the WordNet lexicon.

A formal approach to the identification of a primitive set of words would consider the basic atoms of WordNet extensions (Extended WordNet, OntoWordNet, FrameNet) which represent word definitions in a logical framework. This way, the application of WordNet and its extensions is an important aspect of our proposal for language grounding. However, one further extension is required to incorporate grounded information into WordNet in the direction of a sensorimotor WordNet, designated here as WordNet^{SM}. For the sensorimotor information associated with the basic words, we presented the Human Activity Language as a possible representation in WordNet^{SM}. Despite the possibility of other representations, a linguistic approach takes advantage of the theory of Automatic Speech Recognition and Natural Language Processing.

References

Bailey D., Chang N., Feldman J., and Narayanan S. (1997) Extending Embodied lexical development. In Proceedings of the 20^{th} Annual Meeting of the Cognitive Science Society.

Basili R., Pazienza M., and Velardi, P. (1996) *An empirical symbolic approach to natural language processing*. Artificial Intelligence Journal, 85/1-2, pp. 59–99.

Fellbaum C. (Ed.) (1998) *WordNet: An Electronic lexical Database*. MIT Press, Cambridge, MA.

Gangemi A., Navigli R., and Velardi P. (2003) The OntoWordNet project: extension and acclimatization of conceptual relations in WordNet. In Proceedings of International Conference on Ontologies, Databases and Applications of Semantics.

Gibson J. (1979) *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA.

Guerra-Filho G. (2005) *Optical Motion Capture: Theory and Implementation*. To appear in the Brazilian Computing Society Revista de Informática Teórica e Aplicada.

Guerra-Filho G., Fermüller C., and Aloimonos Y. (2005) Discovering a language for human activity. To appear in Proceedings of AAAI 2005 Fall Symposium: "From Reactive to Anticipatory Cognitive Embodied Systems".

Levin B. (1993) *English Verb Classes and Alternations*. Chicago University Press, Chicago, IL.

Harabagiu S., Miller G., and Moldovan D. (1999) WordNet 2 – A morphologically and semantically enhanced resource. In Proceedings of SIGLEX-99, pp. 1–8.

Harnard S. (1990) *The symbol grounding problem.* Physica, D42, pp. 335–346.

Harrow A. (1972) *A Taxonomy of the Psychomotor Domain: A Guide for Developing Behavioral Objectives*. David McKay Company, Inc., New York, NY.

Joanis E. and Stevenson S. (2003) A General Feature Space for Automatic Verb Classification. In Proceedings of the 10^{th} Conference of the European ACL, pp. 163–170.

Merlo P. and Stevenson S. (2001) *Automatic Verb Classification Based on Statistical Distributions of Argument Structure*. Computational Linguistics, 27/3, pp. 373–408.

Miller G., Beckwith R., Fellbaum C., Gross D., and Miller K. (1990) *Five papers on Wordnet*. International Journal of Lexicography, 3/4.

Miller G. and Fellbaum C. (1991) *Semantic networks of English*. Cognition, 41/1-3, pp. 197–229.

Roy D. (2005) Semiotic schemas: a framework for grounding language in action and perception. To appear in Artificial Intelligence.

Schulte im Walde S. and Brew C. (2002) Inducing German Semantic Verb Classes from Purely Syntatic Subcategorisation Information. In Proceedings of the 40^{th} Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 223–230.

Siskind J. (2001) *Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic*. Journal of Artificial Intelligence Research, 15, pp. 31–90.

Vázquez G., Fernández A., Castellón I., and Martí M. (2000) Clasificación Verbal: Alternancias de Diátesis. Number 3 in Quaderns de Sintagma, Universitat de Lleida.