Argument Selection and Alternations in German VP-Idioms

Christiane Fellbaum

Department of Psychology, Princeton University
and
Berlin-Brandenburgische Akademie der Wissenschaften

Idioms differ not only semantically from the literal language in exhibiting varying degrees of opacity, they also frequently violate the syntactic rules of free language. This talk examines German VP-idioms with respect to their argument selection and alternation behavior.

Both argument reduction and argument augmentation are frequently found in VP-idioms. We address the following questions: Does the idiosyncratic subcategorization behavior of verbs in idioms correlate with their semantic transparency within the VP? Do verbs' idiosyncratic argument selections create frames with their own, lexeme-independent idiomatic meaning?

With respect to argument alternations, we examine to what extent argument alternation patterns that have been identified in the literal language apply to idioms, and whether alternations in idioms are subject to the same semantic constraints. Finally, we examine the specific use idioms make of argument alternations.

Which Verb Classes and Why?

Jean-Pierre Koenig, Gail Mauner, Breton Bienvenue, and Anthony Davis

Linguistics Department
University of Buffalo

Within linguistics, verb classes have often been used to organize lexical knowledge along both syntactic lines (e.g., subcategorization) and semantic lines (e.g., types of denoted semantic relations). The two enterprises converge in so far as subcategorization can be for the most part predicted from the meaning of verbs (cf. Levin (1993) among others) and that valence alternations are for the most part semantically motivated (cf. Pinker (1989) among others). In the first part of this talk I will examine linguistic research that addresses three questions:

1. Under what assumptions is it true that semantic properties determine (for the most part) syntactic subcategorization properties?
2. Why do certain semantic properties, but not others, matter for predicting syntactic classes?
3. Why do we have semantically-based semantic alternations (see Dowty (2001))?

The second part of this talk will examine psycholinguistic research that bears on the existence within the mental lexicon of semantically defined verb classes that have no syntactic reflexes. I will provide experimental evidence for the following points:

1. Because of information-theoretic reasons, class size matters: Smaller semantic classes play a stronger role than all-inclusive semantic classes;
2. (Semantically-defined) verb classes without any syntactic reflexes play a role in processing filler-gap dependencies in reading, in predicting eye-movements in a “visual world” paradigm, and in syntactic priming;
3. Both properties of typical fillers of a semantic role and abstract participant role properties are accessed when a verb’s meaning is accessed.

Computational Experiments on Verb Classes

Paola Merlo

Département de Linguistique
Université de Genève

Recent developments in Natural Language Processing have focussed on data-driven statistical techniques and exploited existing text as a learning resource. Learning methodologies have been developed based on the assumption that fundamental linguistic notions, such as those defining verb semantics and verb classes, can be automatically learned by appropriately developed statistics on a large corpus.

I will draw on many experiments on verb classification and disambiguation to illustrate what learning features and techniques have been found to work and which do not give satisfactory results to date. We will see that cross-linguistic transfer effects are very helpful learning cues in classification, while appropriate and successful use of alternations still eludes us.

Representing verb transitivity information: Evidence from syntactic priming

Manabu Arai
Department of Psychology
University of Dundee
Dundee, Scotland, DD1 4HN
m.arai@dundee.ac.uk

Roger P.G. van Gompel
Department of Psychology
University of Dundee
Dundee, Scotland, DD1 4HN
r.p.g.vangompel@dundee.ac.uk

Jamie Pearson
Department of Psychology
University of Edinburgh
Edinburgh, Scotland, EH8 9JZ
jamie.pearson@ed.ac.uk

1 Introduction

It is a well-known phenomenon that language users tend to repeat syntactic structures across utterances. This phenomenon occurs even when there are no lexical, semantic or prosodic relations between two consecutive sentences (Pickering & Branigan, 1999). This tendency is widely referred to as syntactic priming (e.g., Bock, 1986, 1989; Bock & Loebell, 1990). It has been shown that syntactic priming can be used to investigate the representation of syntactic structures (e.g., Pickering & Branigan, 1998, Branigan, Pickering & Cleland, 2000). Pickering and Branigan (1998) investigated syntactic priming of two alternative ditransitive structures (double object, DO structures, e.g., *the racing driver showed the team manager the torn overall* and prepositional object, PO structures, e.g., *the racing driver showed the torn overall to the team manager*) using a written completion task. They found that participants tended to complete target fragments using the same structure as they had used in the preceding prime sentences. They observed this tendency when the verb in the target fragment was different from the one in the prime as well as when the verb in the target was the same as in the prime. Therefore, they argued that information about the way a verb combines with other linguistic constituents (they call this 'combinatorial information') is shared between verbs. However, they observed a stronger priming effect when the verb in the prime and target was repeated. Pickering and Branigan (1998) accounted for these results with a model developed from Roelofs' (1992, 1993) lexical network model. They argued that combinatorial information is directly linked to each verb at the lemma stratum. For example, when the prime verb *show* is used with the PO construction, it activates not only the combinatorial PO node but also the link between this node and the particular verb *show*. Because of the residual activation of this link, people show a stronger tendency to produce PO completions when they see the same verb show in the target compared to when they see a different verb such as *give.*

Pickering and Branigan's model assumes that combinatorial information is directly linked to each individual verb. We refer to this form of representation as *lexically specific* because syntactic information is associated with each individual verb. This explains why priming in PO/DO structures is stronger when the verb is repeated than when it is not. This type of representation can be distinguished from a *category-general representation*, that is, combinatorial information is represented independently from lexical information and syntactic information is associated with a word class (e.g., verbs) as whole. If syntactic information is represented at the category-general level, priming should be no stronger when the verb is repeated than when it is not (in contrast with Pickering and Branigan's findings from PO/DO structures).

The aim of the present experiments was to investigate the representation of verbs' transitivity information. Until now, the representation of transitivity information has not been investigated systematically. Although it is generally assumed that people represent transitivity information and use it during sentence processing (e.g., Clifton, Frazier & Connine, 1984; Stowe, Tanenhaus, and Carlson, 1991), it is unclear exactly how transitivity information is stored.

2 Experiment 1

Experiment 1 investigated how monotransitive and intransitive structures are represented using the syntactic priming method. In order to address the question of whether monotransitivity and intransitivity information are represented at the lexically-specific or category-general level, we investigated whether priming is stronger when the verb in the prime and target is repeated than when it is not.

2.1 Method

We adopted a spoken sentence completion task (e.g. Branigan et al., 2000). Participants read one of the prime sentences in (1) aloud and next they read the target fragment (2) aloud and completed it. We used verbs that can be used as either monotransitives or intransitives.

Prime sentences were either monotransitive sentences (1a and c) or intransitive sentences (1b and d). We also manipulated whether the verb in the prime sentence was repeated in a target (1a and b) or not repeated (1c and d). The target fragment could be completed as either a monotransitive completion (e.g., *While the prisoner was bullying the inmate, he was put into detention*) or an intransitive completion (e.g., *While the prisoner was bullying, other prisoners joined in*).

1a. The teenager was bullying the man. (monotransitive prime, verb repeated)
1b. The teenager and the man were bullying. (intransitive prime, verb repeated)
1c. The teenager was jeering the man. (monotransitive prime, verb not repeated)
1d. The teenager and the man were jeering. (intransitive prime, verb not repeated)

2. While the prisoner was bullying….... (target)

2.2 Results

The percentage of monotransitive completions out of the total number of monotransitive and intransitive completions was taken as a measure of the activation of the monotransitive structure.

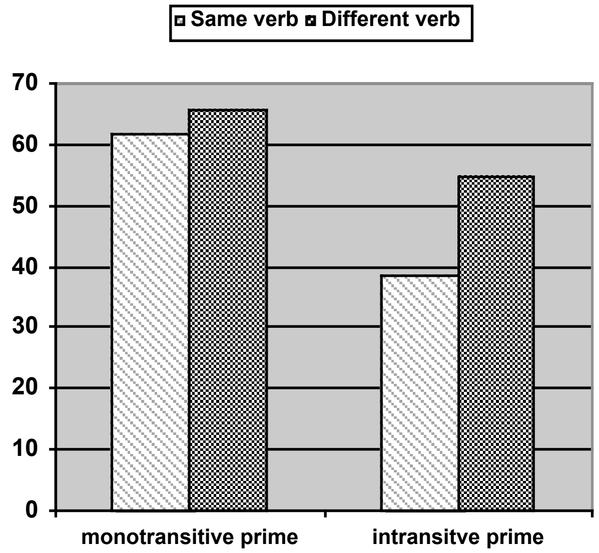


Figure 1: Percentage of monotransitive completions out of all mono- and intransitive completions

Figure 1 shows the percentage of monotransitive completions. The results showed that participants produced more monotransitive completions to (2) after reading monotransitive (1a and c) than intransitive primes (1b and d). Furthermore, there was an interaction between prime structure and verb repetition. After intransitive primes, participants produced fewer monotransitive completions, in other words, more intransitive completions when the verb in prime and target was the same than when the verbs were different. Most interestingly, however, we did not observe stronger priming after monotransitive primes when the verb was repeated. That is, there was a verb repetition effect for the intransitive conditions but not for the monotransitive conditions.

2.3 Conclusions

The finding that priming for intransitives was stronger when the verb was repeated than when it was not suggests that the representation of intransitivity information is directly associated with each individual verb. In contrast, priming for monotransitives did not differ depending on whether the verb was repeated or not. This suggest that the representation of monotransitivity information is not directly associated with a particular verb, but rather associated with the class of verbs as a whole. In other words, intransitivity information is represented at the lexically-specific level, whereas monotransitivity information is represented at the category-general level.

3 Experiment 2

Experiment 2 addressed the question of whether the absence of a verb repetition effect for monotransitives is unique to the particular verbs used in Experiment 1, that is, verbs that can be used either as monotransitives or intransitives. In Experiment 2, we used verbs that can be used either as monotransitives or ditransitives and investigated how transitivity information is represented for these verbs.

3.1 Method

The method was the same as in Experiment 1. Prime sentences were either monotransitive sentences (3a and c) or ditransitive sentences (3b and d). In (3a and b), the verb in a prime sentence was repeated in the target whereas it was not in (3c and d). The target fragment could be completed as either a monotransitive completion (e.g., *The uncle sold his watch*) or a ditransitive completion (e.g., *The uncle sold his brother the car*).

3a. The performer sold the ticket. (monotransitive prime, verb repeated)
3b. The performer sold the tourist the ticket. (ditransitive prime, verb repeated)
3c. The performer offered the ticket. (monotransitive prime, verb not repeated)
3d. The performer offered the tourist the ticket. (ditransitive prime, verb not repeated)

4. The uncle sold...…. (target)

3.2 Results

The percentage of monotransitive completions out of the total number of monotransitive and ditransitive completions was taken as a measure of the activation of the monotransitive structure.

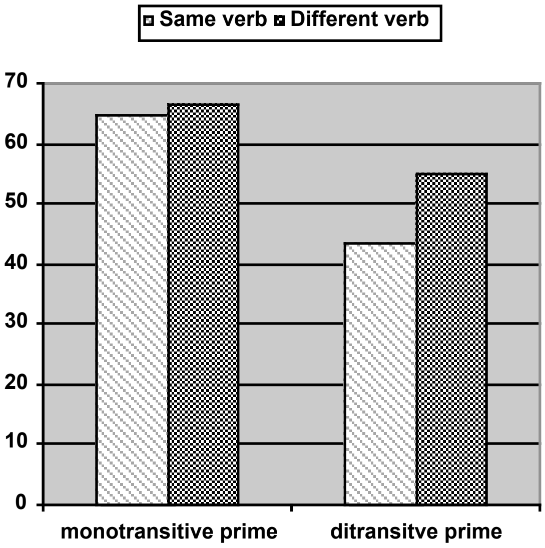


Figure 2: Percentage of monotransitive completions out of all mono- and ditransitive completions

Figure 2 shows the percentage of monotransitive completions. The results showed that participants produced more monotransitive completions to (4) after monotransitive primes (3a and c) than after ditransitive primes (3b and d). Most importantly, we observed an interaction between prime structure and verb repetition. After ditransitive primes, we observed a stronger priming effect when the verb was repeated than it was not, whereas there was no evidence that the priming effect for monotransitives was stronger when the verb was repeated than when it was not. That is, there was a verb repetition effect for ditransitives but not for monotransitives.

3.3 Conclusions

Experiment 2 showed that mono/ditransitive verbs represent monotransitivity information in the same way as mono/intransitive verbs do. The results suggested that monotransitivity information is represented at the category-general level (confirming the conclusions from Experiment 1) whereas the representation of ditransitivity information is lexically specific, similar to intransitivity information investigated in Experiment 1.

4 Experiment 3

In Experiments 1 and 2, there was no evidence that the priming effect for monotransitives was stronger when the verb was repeated between the prime and target than when it was not. One possibility is that monotransitives do not prime in either the verb repeated or non-repeated conditions, which would result in no verb repetition effect. The monotransitive structure is a highly frequent structure, and therefore, the activation of this structure can perhaps not be boosted any further by a monotransitive prime sentence. We tested this hypothesis by investigating whether monotransitives (5a) prime relative to a baseline condition (5b) that did not contain a verb. We also included an intransitive prime condition (5c).

4.1 Method

The method was the same as in Experiments 1 and 2. After both the monotransitive (5a) and intransitive primes (5b), the verb was repeated in the target. The baseline sentences (5c) were composed of adverbs and adjectives and did not contain verbs and nouns, as they may be associated with combinatorial information (cf. Cleland & Pickering, 2003; Pickering & Branigan, 1998) and may prime.

5a. The ambulance driver overtook the policewoman. (monotransitive prime)
5b. The ambulance driver and the policewoman overtook. (intransitive prime)
5c. Always passionate and very creative. (baseline prime)

6. When the motor cyclist overtook……. (target)

4.2 Results

The percentage of monotransitive completions out of the total number of monotransitive and intransitive completions was taken as a measure of the activation of the monotransitive structure.

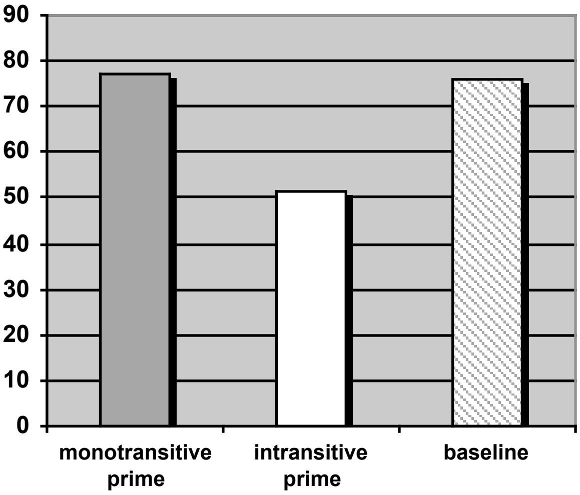


Figure 3: Percentage of monotransitive completions out of all mono- and ditransitive completions

Figure 3 shows the percentage of monotransitive completions. The results showed that participants produced more intransitive completions after intransitive primes than after monotransitive primes and baseline sentences. Most importantly, they did not produce more monotransitive completions after monotransitive primes than after baseline sentences. Therefore, there was no priming effect for monotransitives.

4.3 Conclusions

The results of Experiment 3 showed that the intransitive structure primes whereas the monotransitive structure does not. It might be argued that the null priming effect for monotransitives is due to a ceiling effect. However, this account is unlikely as the percentage of monotransitive completions is well below 100%. We discuss reasons for the absence of a priming effect for monotransitive structures below.

5 Discussion

In Experiment 1, we observed that participants produced more monotransitive completions after reading monotransitive primes than after reading intransitive primes and more intransitive completions after reading intransitive primes than after reading monotransitive primes. Most importantly, priming of intransitives was stronger when the verb was repeated than when it was not, whereas priming of monotransitives was the same, regardless of whether the verb was repeated. Experiment 2 showed that participants produced more ditransitive completions after reading ditransitive primes than after reading monotransitive primes and the ditransitive priming was larger when the verb was repeated than when it was not. However, as in Experiment 1, priming for monotransitives was no stronger when the verb was repeated than when it was not. In Experiment 3 , we replicated Experiment 1, observing that participants produced fewer monotransitive completions after reading intransitive primes than after reading monotransitive primes and, furthermore, they produced fewer monotransitive completions after reading intransitive primes than after reading baseline sentences. Most importantly, Experiment 3 showed that participants produced no more monotransitive completions after monotransitives than after baseline sentences, indicating that monotransitives did not prime at all. This suggests that the absence of a verb repetition effect for monotransitives in Experiments 1 and 2 was due to a general absence of priming from monotransitives.

How can the absence of monotransitive priming be explained? First, people may not represent the monotransitive structure. However, this is implausible, because without such a representation, the language processor would not know how to process monotransitive sentences and people would not be able to determine that monotransitive sentences are grammatical. A much more plausible alternative account is that the level of activation for monotransitives is already at a maximum level and cannot be boosted any further. There are two ways that the monotransitive structure can have a maximum level of activation. One possibility is that for some individual verbs, the monotransitive structure is maximally activated because they are nearly always used as monotransitives and therefore cannot be activated any further. This is a *lexically-specific maximum activation* because the maximum activation only occurs for verbs that have an extremely strong monotransitive bias, but not for others. However, it is somewhat unlikely that this explains the lack of priming in our experiments, because the verbs that we used in our experiments were not particularly strongly biased towards the monotransitive structure. The second, more plausible explanation is that the monotransitive structure is maximally activated for the class of verbs in general. Across all verbs, monotransitives are much more frequent than either intransitives or ditransitives, so it seems plausible to assume that the monotransitive structure has a *category-general maximum activation*.

We believe that a category-general maximum activation is also the most economical way of representing transitivity information. Language users very frequently hear or read monotransitive structures, so keeping track of all occurrences of monotransitives is relatively costly. In contrast, intransitive and ditransitive structures occur less frequently, so keeping track of their activation is relatively easy. Furthermore, because almost all verbs can be used as monotransitives (even so-called intransitive verbs such as *sneeze*: *Peter sneezed blood*), there is no need to represent monotransitivity information for individual verbs, especially given that impossible monotransitive sentences can often be ruled out by their semantics (e.g., **The man sneezed the nose*). In other words, semantic information is often sufficient to determine whether a monotransitive sentence is acceptable.

Finally, the idea that monotransitives constitute the default structure is also consistent with the findings from child language. There is evidence that with age, children show an increasing tendency to overgeneralize the monotransitive use of intransitive verbs (e.g., **Peter giggled me* or **she cried her*), whereas the tendency to overgeneralize the intransitive use of intransitive verbs decreases with age (e.g., **John hits*) (Brooks & Tomasello, 1999; Maratsos, Gudeman, Gerard-Ngo, & Dettart, 1987).

The fact that children produce more monotransitive overgeneralizations with age suggests that they gradually come to assume that any verb can be used as monotransitive. In contrast, the fact that children do not overgeneralize intransitive usage suggests that they gradually come to assume that not all verbs can be used as intransitive. Therefore, as children grow older, the monotransitive structure becomes the default structure and is represented at the category-general level, whereas intransitives are represented at the lexically-specific level.

Our results can be integrated into the Pickering and Branigan's (1998) network model by extending it to the representation of monotransitivity and intransitivity information. As mentioned in Introduction, their model assumes that combinatorial information is associated with each individual verb node. More specifically, combinatorial information about the PO and DO structure is directly associated with each individual verb node. Priming in the repeated verb conditions is stronger than priming in the different verb conditions because the activation of the link between a particular verb node and the DO or PO node is boosted. Our experiments showed that intransitivity information is also associated with each individual verb, in the same way as information about the DO and PO structure. In contrast, we have argued that monotransitivity information is associated at the categorically-general level, and is therefore not directly associated with each individual verb. We assume that combinatorial information of the monotransitive structure is associated with the verb-category node and is therefore represented for the class of verbs as a whole. As assumed by Pickering and Branigan (1998), the verb-category node is shared among all verbs and because it is activated whenever any verb is used, the node is *inherently activated*, that is, it has the maximum level of activation. The absence of a priming effect for monotransitives suggests that the link between this category-general node and monotransitive combinatorial information also has a maximum level of activation. This contradicts many current linguistic theories that assume that each lexical item incorporates syntactic information on what type of argument structures it can take (Chomsky, 1981; Pollard & Sag, 1994). Instead, we claim that not all transitivity information is associated with individual verbs: intransitivity and ditransitivity information is associated with individual verbs, but monotransitivity information is not.

References

Bock, J.K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, **18**, 355-387.

Bock, K. (1989). Closed-Class Immanence in Sentence Production. *Cognition,* **31,** 163-186.

Bock, K., & Loebell, H. (1990). Framing Sentences. *Cognition*, 35, 1-39.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition,* **75,** B13-B25.

Branigan, H.P., Pickering, M.J., Stewart, A.J., & McLean, J.F. (2000). Syntactic priming in spoken production: Linguistic and temporal interference. *Memory & Cognition*, **28**, 1297-1302.

Brooks, P. J., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language,* **75,** 720-738.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, The Netherlands: Foris.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language,* **49,** 214-230.

Clifton, C., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, **23**, 696-708.

Maratsos, M., Gudeman, R., Gerard-Ngo, P., & Dettart, G. (1987). A study of novel word learning: The productivity of the causative. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 89-113). Hillsdale, NJ: Erlbaum.

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences,* **3,** 136-141.

Pickering, M.J., & Branigan, H.P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, **39**, 633-651.

Pollard, C., & Sag, I.A. (1994) *Head-driven phrase structure grammar*. Stanford, CA & Chicago, IL: CSLI & University of Chicago Press.

Roelofs, A. (1992). A Spreading-Activation Theory of Lemma Retrieval in Speaking. *Cognition,* **42,** 107-142.

Roelofs, A. (1993). Testing a Non-Decompositional Theory of Lemma Retrieval in Speaking - Retrieval of Verbs. *Cognition,* **47,** 59-87.

Stowe, L.A., Tanenhaus, M.K., & Carlson, G. (1991). Filling gaps on-line: Use of lexical and semantic information in sentence processing. *Language and Speech*, **34**, 319-340.

On the Semantic Determinants of Inflection Class Membership: Evidence from Lithuanian

Peter M. ARKADIEV
Department of Typology and Comparative Linguistics,
Institute for Slavic Studies, Russian Academy of Sciences,
Leninski Avenue, 32-A
Moscow, Russian Federation, 117334
alpgurev@yandex.ru

Abstract

In this paper I argue that inflection class membership among the so-called 'primary' verbs in Lithuanian, which has always been considered to be extremely idiosyncratic, is at least partly predictable from the verb's semantics. The most important semantic parameters responsible for inflection class assignment are agentivity of the verb's highest ranking participant (thus most transitive and agentive intransitive 'primary' verbs share the same morphological features whereas non-agentive intransitives fall into another inflectional class) and the inherent aspectual properties of the verb (intransitive verbs denoting atelic and telic processes fall into different classes). These semantic features are cross-linguistically recognized as relevant for 'unaccusativity' or 'split intransitivity'; thus Lithuanian inflectional morphology may be subsumed under a typologically well-established pattern.

1 Introduction

The verbal system of Lithuanian is notorious for both number and complexity of various morpho(phono)logical features whose combinations produce quite a large inventory of inflectional classes; see (Dressler et al., 2004) for a comprehensive analysis. The greatest diversity of patterns shows itself with the so-called 'primary' verbs (those whose infinitive is formed by attaching the suffix *-ti* directly to the root, like *bėg-ti* 'to run') which distinguish about 15 distinct patterns, see Table 1 for only a small subset of actual possibilities. The attempts to account for the distribution of these patterns in phonological or morphophonological terms (see e. g. Ambrazas (ed.), 1997) turn out to be inadequate, especially when trying to predict whether the verb would fall into one of the two largest subclasses of 'primary' verbs: those having the nasal infix or the suffix *-st-* in the Present stem[1] (e. g. *migti* 'to fall asleep', *dingti* 'to disappear' in Table 1; they will be called *n/st*-verbs hereafter) vs. those palatalizing the last consonant of both Present and Past stems[2] (e.g. *gerti* 'to drink' in Table 1, *knarkti* 'to snore'; they will be called *j*-verbs in the subsequent text).

Infinitive	Present3Sg	Past3Sg	Gloss
bėgti	*bėga*	*bėgo*	'run'
migti	*minga*	*migo*	'fall asleep'
dingti	*dingsta*	*dingo*	'disappear'
gimti	*gimsta*	*gimė*	'be born'
gerti	*geria*	*gėre*	'drink'

Table 1. Some inflectional classes of Lithuanian primary verbs

The main goal of this paper is to argue that the verb's assignment to one of the two major subclasses (viz. the aforementioned *n/st*-class and *j*-class) is to a great extent determined by its semantics.

2 The semantics of *n/st*-verbs

A closer examination of the meanings of verbs belonging to the *n/st*-class reveals that with minor exceptions they form a semantically coherent class: almost 90% of these verbs (the class comprises more that 250 lexemes) denote telic eventualities whose only participant is a Patient (viz., has enough Proto-Patient properties in the sense of (Dowty, 1991), (Ackerman and Moore, 2001)): *aušti* 'to cool down', *blukti* 'to fade away', *dužti* 'to break (intr.)', *gižti* 'to turn sour', *kimti* 'to become hoarse', *lipti* 'to stick', *pigti* 'to become

[1] The distribution of the infix and the suffix themselves is purely phonological, see (Stang, 1942).

[2] Palatalization is orthographically expressed by -*i*- between the consonant and the following vowel or by the ending -*ė* in the Past forms.

cheaper', *rausti* 'to become red' etc. These verbs may be characterized as denoting externally caused eventualities in the sense of (Levin and Rappaport Hovav, 1995, 1998) thus sharing the following lexico-semantic representation:

(1) [*ACTIVITY*] CAUSE [BECOME[*STATE*(x)]]

The main feature distinguishing these verbs from their transitive counterparts, which often belong to the *j*-class (cf. *linkti* 'to bend (intr.)' — Present *linksta* vs. *lenkti* 'to bend (tr.)' — Present *lenkia*) is that the latter require the explicit specification of both the activity and its instigator, the Agent, while the former leave this semantic component and its participant completely unspecified. Thus the following may serve as refined lexico-semantic representations of *linkti* and *lenkti*:

(2) *linkti*: λy∃x [ACT(x)]CAUSE
[BECOME[*BENT*(y)]]

(3) *lenkti*: λyλx [ACT(x)]CAUSE
[BECOME[*BENT* (y)]]

There are some verbs in the *n/st*-class which at first sight do not conform to the above stated prototype. Those are e.g. agentive[3] intransitives *kilti* 'to rise' and *sprukti* 'to flee' and transitive *justi* 'to (come to) feel' and *mėgti* 'to (come to) like'. However, I believe that at least these putative exceptions can be subsumed under the semantic prototype of the *n/st*-class. The first two verbs denote directed motion and are telic; they have the following lexico-semantic representation:

(4) λx [ACT(x)] CAUSE [BECOME[*STATE*(x)]]

The other pair, although syntactically transitive, are non-canonical dyadic predicates (see e.g. (Tsunoda, 1985) for a cross-linguistic survey of such verbs), whose highest ranking participant has just a few of the Proto-Agent properties; what they have in common with the prototypical telic intransitives is that the change-of-state component embedded into their meaning is predicated of the highest ranking participant (= syntactic subject); cf. similar observations made for auxiliary selection in Dutch in (Lieber and Baayen, 1997). Therefore, it is possible to speculate that inflection class assignment and argument selection in Lithuanian are sensitive to different semantic properties of predicates, but I am not going to pursue this topic further, since I have not investigated it in sufficient depth.

Thus, although not all verbs belonging to the *n/st*-class may be fully subsumed under the semantic prototype of telic patientive intransitives, the class itself may be adequately characterized semantically.

3 The semantic classes of *j*-verbs

The *j*-class is much less semantically homogenous than the *n/st*-class. It comprises almost 400 lexemes of which more than 50 % are (canonical) transitives, such as *verpti* 'to spin (thread)', *arti* 'to plough', *drožti* 'to plane', *ližti* 'to lick', *rėžti* 'to cut', *blokšti* 'to throw', *klausti* 'to ask' etc. The intransitive *j*-verbs form a large group and fall into several subclasses:

(i) verbs of internally caused sound emission: *bimbti* 'to buzz', *gergžti* 'to talk hoarsely', *knarkti* 'croak', *pipti* 'peep' etc.;

(ii) verbs of light or smell emission: *pliksti* 'to shine', *dvokti* 'to stink';

(iii) agentive verbs of manner of motion: *plaukti* 'to swim', *kuisti* 'to run very fast', *lėkti* 'to fly' etc.;

(iv) verbs denoting natural activities, most probably conceptualized as caused by an Agent-like natural force: *bliaukti* 'to flow (of a stream)', *dumti* 'to blow (of the wind)' etc.;

(v) verbs denoting activities with a human protagonist: *brūzti* 'to toil', *žaisti* 'to play' etc.

It is clear that the intransitive *j*-verbs share an important semantic feature: they denote internally caused atelic eventualities. This may be clearly seen from the contrast between agentive verbs of motion belonging to the *j*-class and to the *n/st*-class: the latter are verbs of directed motion (telic) while the former are verbs of manner of motion (atelic), cf. (Levin and Rappaport Hovav, 1990, 1995). The common lexico-semantic representation of intransitive *j*-verbs is the following:

(5) λx [$\text{ACT}_{<\mathit{MANNER}>}$ (x)]

It is also not surprising that both agentive intransitive and transitive verbs fall into the *j*-class: the feature they share is the Activity component predicated of their highest ranking participant, cf. (3) and (5).

4 Other verb classes

Other subclasses of Lithuanian primary verbs have considerably fewer members, and it

[3] However, *kilti* may be used with a whole variety of subjects, not necessarily animate and agentive, cf. *vandens lygis kyla* 'the water level rises' ; besides, like quite a number of non-agentive *n/st* verbs, *kilti* has a transitive *j*-counterpart : *kelti*.

is hard to postulate a coherent semantic basis for any of them. Among the verbs which fall into these minor classes there are both transitive and intransitive predicates, and the latter may be either agentive or patientive.

However, while it is not possible to semantically motivate inflectional properties of *all* members of the minor morphological classes, it seems that such a motivation nevertheless can be found for some such verbs. For instance, consider the following lexemes: *sėsti* ‘to sit down’, *lipti* ‘to climb’, *lįsti* ‘to penetrate into smth.’; they have neither infix/*st*-suffix, nor *j*-suffix: Present *sėda*, Past *sėdo*. What they have in common semantically, as it seems, is both genuine agentivity of the subject (these verbs usually allow only animate subjects) and the ‘change of state’ component. Thus, they do not fall under either prototype stated above, and this is, probably, the reason why they are not assigned to either of the major inflectional classes.

Another small set of predicates for which a putative explanation of their inflectional class membership can be adduced are three labile verbs, which have both causative and inchoative (Haspelmath, 1993) uses: *degti* ‘to burn’, *kepti* ‘to bake’, *virti* ‘to boil’. They belong to yet another small and semantically heterogeneous inflectional class, sharing with the *j*-verbs the Past stem, but lacking any affix in the Present stem: Present *dega*, Past *degė*. Since these verbs conform to both prototypes in their different senses, which fail to be formally differentiated (unlike such pairs as *linkti*/*lenkti* ‘to bend (intr/tr)’), it is not very surprising that they belong to a morphological type distinct from those of canonical transitives and patientive intransitives. It is probably possible to consider their morphological properties as ‘iconically’ reflecting their ‘dual’ semantico-syntactic behaviour: ordinary transitive verbs have *j*-suffix in both stems, while labile verbs palatalize only the Past stem.

Notwithstanding possible semantic motivations for some members of the minor inflectional classes of Lithuanian ‘primary’ verbs, I believe that only the major classes, namely the *n/st*-class and the *j*-class, can be unequivocally characterized semantically.

5 Interim summary

In the preceding sections I have tried to show that inflectional class assignment with ‘primary’ verbs in Lithuanian is motivated by the semantic structure of these lexical items. The correlation between semantic features and inflection class may be seen in Table 2 (based on a list of ‘primary’ verbs with consonant-final roots taken from (Lyberis, 1962)). As the figures indicate, there is a statistically highly significant interdependency between semantic and morphological classes of ‘primary’ verbs in Lithuanian (especially with monadic verbs); moreover, it is possible to pin down single components of meaning responsible for inflectional class assignment:

(6) BECOME[*STATE*(x)] → *n/st*-class.

(7) ACT(x) → *j*-class.

	j	*n/st*	Other	Total
Transitive	**247**	8	51	306
Agentive intransitive	**121**	7	7	135
Patientive intransitive	7	**237**	4	248
Total	375	252	62	689

Table 2: The distribution of semantic and morphological classes of Lithuanian ‘primary’ verbs

If both components co-occur in the lexico-semantic representation of a verb and are predicated of the same participant, the conflict is resolved either by some sort of hierarchical ranking of these parameters (thus, for *kilti* ‘to rise’, which belongs to the *n/st*-class, the ranking is (6) > (7)) or by assigning the verb to some minor inflectional class (e.g., agentive telic *sėsti* ‘to sit down’ has neither palatalized stem-final consonant nor infix or suffix). Such variation is not unexpected, since it is in the non-prototypical cases that the least language-internal and cross-linguistic consistency of patterns usually shows up.

Thus, it is possible to conclude that inflection class membership among Lithuanian ‘primary’ verbs, especially in their intransitive subset, has a clear, although not a 100 %, semantic motivation.

6 Typological perspective: Georgian

In order to see that the phenomena discussed above are not merely an idiosyncrasy of a language with highly irregular inflectional morphology, let us briefly look at the data from an unrelated language with strikingly similar matches between lexical semantics and verbal morphosyntax, namely Georgian.

As is widely acknowledged (see (Vogt, 1971), (Harris, 1981), (Holisky, 1979, 1981), (Merlan, 1985), (Van Valin, 1990) for both descriptive generalizations and explanatory proposals), there are three major productive classes of verbs in Georgian, all of which are more or less homogenously semantically motivated. The morphosyntactic properties of Georgian verbal classes are summarized in Table 3; they include subject agreement morphology (here are relevant only 3SgPresent, 3PlPresent, and 3PlAorist suffixes) and case assignment to subject and object in the Aorist tense.

Class	Case-marking	Agreement
I	Sb: Erg — Ob: Nom	*-s* — *-en* — *-es*
II	Sb: Nom	*-a* — *-an* — *-nen*
III	Sb: Erg	*-s* — *-en* — *-es*

Table 3. Verb classes in Georgian

Semantic properties of the verbs belonging to these classes may be outlined as follows (see (Harris, 1981) and (Holisky, 1981) for an extensive treatment; I consider only underived verbs):

Class I contains transitive (dyadic) verbs: *mok'lavs* 'to kill', *dac'ers* 'to write', *dagvis* 'to sweep smth out', *šek'eravs* 'to sew', *micems* 'to give' etc.

Class II mainly contains verbs denoting telic eventualities, among which are both patientive and agentive: *mok'vdeba* 'to die', *darčeba* 'to remain', *dadneba* 'to melt', *dadgeba* 'to stand up' etc.

Class III contains verbs denoting atelic eventualities; the range of meanings possible with these verbs resembles very much that of Lithuanian intransitive *j*-verbs:

(i) verbs of sound emission: *bzuk'unebs* 'to buzz', *laklakebs* 'to chat', *xorxocebes* 'to laugh loudly' etc.;

(ii) verbs of light emission: *bdγrialebs* 'to glisten', *varvarebs* 'to flare', *rialebs* 'to twinkle' etc.

(iii) verbs denoting 'motion without displacement': *babanebs* 'to tremble', *trtis* 'to shake' etc.;

(iv) verbs denoting non-directed motion: *goravs* 'to roll', *xt'is* 'to jump', *curavs* 'to swim', *parpatebs* 'to flit' etc.;

(v) verbs denoting natural processes: *grgvinavs* 'to thunder', *tovs* 'to snow', *kris* 'to blow (of the wind)' etc.

(vi) verbs denoting activities with a human protagonist: *tamašobs* 'to play', *mušaobs* 'to work', *cek'vavs* 'to dance' etc.

Thus, verb classes in Georgian have well-grounded semantic motivation, which, moreover, is quite similar to that of Lithuanian *j*- and *n/st*-verbs. Besides, just as Lithuanian atelic verbs pattern with transitive verbs morphologically, so do their Georgian counterparts: it is evident from Table 3 above that classes I and III share agreement morphemes (however, these verbs are dissimilar in other important morphological respects).

This evident similarity in the semantic properties of verbal classes in two unrelated languages which have never been in any contact cannot be accidental and must be motivated by cross-linguistically valid or even universal patterns linking lexical semantics, argument structure and morphosyntax (see (Lazard, 1985), (Van Valin, 1990), (Verhaar, 1990), (Mithun, 1991), (Levin and Rappaport Hovav, 1995), (Kibrik, 1997), (Croft, 1998), (Alexiadou et. al (eds.), 2004) for various attempts at explaining such and similar cross-linguistic similarities).

7 Summary and conclusions

In this paper I hope to have shown that all idiosyncrasies notwithstanding, it is possible to arrive at a fairly reliable predictability of inflection class of a Lithuanian 'primary' verb on the basis of its lexical semantics. Certainly, there is no exact 100 % matching between semantic features and morphological properties, but the correlation is nevertheless statistically highly significant.

Having compared Lithuanian data with that of a well-studied language, Georgian, I have argued that there is a striking and undoubtedly non-accidental similarity between verbal classes in these languages. Certainly, the Georgian verbal system is much more semantically transparent than that of Lithuanian; however, the verbal lexicon of both languages seems to be structured by the same semantic features, viz. agentivity/patientivity and telicity/atelicity.

What is also important to mention is the fact that the semantic parameters of inflection class assignment of intransitive verbs in Lithuanian and Georgian coincide with those usually regarded as determining the unaccusative vs. unergative classification of verbs, cf. (Van Valin, 1991), (Levin, Rappaport Hovav, 1995). Actually, with respect to Georgian it was argued by Harris (1981, 1982) on the basis of syntactic behaviour (e.g., case marking of subjects) of verbs of Classes II and III, that the former are unaccusative, while the latter are unergative. While it will require further investigations to determine whether Lithuanian intransitive *j*-verbs are syntactically unergative, and *n/st*-verbs unaccusative (see (Timberlake,

1982) for attempts to discover unaccusative diagnostics for Lithuanian), it is already significant that morphological properties of Lithuanian verbs conform to typologically well-established patterns.

8 Acknowledgements

I am deeply grateful to Alexej V. Andronov, Anna V. Dybo, Leonid I. Kulikov, Elena V. Paducheva, Barbara H. Partee, Vladimir A. Plungian, Yakov G. Testelec and Vladimir N. Toporov for various kinds of help which I received from them in the course of the investigation on which the present paper is based, and also to two anonymous reviewers for useful suggestions. All faults and shortcomings are mine.

References

F. Ackerman and J. Moore. 2001. *Proto-Properties and Grammatical Encoding. A Correspondence Theory of Argument Selection*. Stanford: CSLI.

A. Alexiadou, E. Anagnostopoulou, M. Everaert (eds.), *The Unaccusativity Puzzle*. Oxford: Oxford University Press, 2004.

V. Ambrazas (ed.). 1997. *Lithuanian Grammar*. Vilnius: Baltos Lankos.

W. C. Croft. Event structure in argument linking. In *The Projection of Arguments. Lexical and Compositional Factors*. P. 21 — 63. M. Butt, W. Geuder (eds.), Stanford: CSLI Publications.

D. R. Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67-3: 547 — 619.

W. U. Dressler, M. Kilani-Schoch, L. Pestal, N. Gagarina and M. Pöchträger. 2004. On the typology of inflection class systems. Paper presented at the 11th International Morphology Meeting, Vienna.

A. C. Harris. 1981. *Georgian Syntax. A Study in Relational Grammar*. Cambridge: Cambridge University Press.

A. C. Harris. 1982. Georgian and the unaccusative hypothesis. *Language*, 58-2, 290 — 306.

M. Haspelmath. 1993. More on the typology of inchoative/causative verb alternations. In *Causatives and Transitivity*. P. 87 — 120. B. Comrie, M. Polinsky (eds.), Amsterdam, Philadelphia: John Benjamins.

D. A. Holisky. 1979. On lexical aspect and verb classes in Georgian. In *The Elements: A Parasession on Linguistic Units and Levels, Including Papers from the Conference on Non-Slavic Languages of the USSR (The 15th Annual Meeting of the Chicago Linguistic Society)* P. 390 — 401. P. R. Clyne, W. F. Hanks, C. L. Hofbauer (eds.), Chicago: Chicago Linguistic Society.

D. A. Holisky. 1981. *Aspect and Georgian Medial Verbs*. New York: Caravan.

A. E. Kibrik. 1997. Beyond subject and object: Toward a comprehensive relational typology. *Linguistic Typology*, 1-3, 279 — 346.

G. Lazard. 1985. Anti-impersonal verbs, transitivity continuum and the notion of transitivity. In *Language Invariants and Mental Operations*. P. 115 — 123. H. Seiler, G. Brettschneider (eds.), Tübingen: Narr.

B. Levin and M. Rappaport Hovav. 1990. The lexical semantics of verbs of motion: The perspective from unaccusativity. In *Thematic Structure: Its Role in Grammar*. P. 247 — 269. M. Roca (ed.), Berlin, New York: Mouton de Gruyter.

B. Levin and M. Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge (Mass.): MIT Press.

B. Levin and M. Rappaport Hovav. 1998. Building verb meanings. In *The Projection of Arguments. Lexical and Compositional Factors*. P. 97 — 134. M. Butt, W. Geuder (eds.), Stanford: CSLI.

R. Lieber and H. Baayen. 1997. A semantic principle of auxiliary selection in Dutch. *Natural Language and Linguistic Theory*, 15-4: 789 — 845.

A. Lyberis. 1962. *Lietuvių-rusų kalbų žodynas*. Vilnius.

F. Merlan. 1985. Split intransitivity: Functional oppositions in intransitive inflection. In *Grammar Inside and Outside the Clause. Approaches to Theory from the Field*. P. 324 — 362. J. Nichols, A. Woodbury (eds.), Cambridge: Cambridge University Press.

M. Mithun. 1991. Active/agentive case marking and its motivations. *Language*, 67-3, 510 — 546.

Chr. S. Stang. 1942. Das slavische und baltische Verbum. *Skrifter utgitt av Det Norske Videnskaps-Akademi i Oslo. II. Hist.-Filos. Klasse,* No. I, 1 — 280.

A. Timberlake. 1982. The impersonal passive in Lithuanian. In *Proceedings of the 8th Annual Meeting of the Berkeley Linguistics Society*. P. 508 — 524, Berkeley: Berkeley Linguistics Society.

T. Tsunoda. 1985. Remarks on transitivity. *Journal of Linguistics*, 21-2, 385 — 396.

J. M. Verhaar. 1990. How transitive is intransitive? *Studies in Language*, 14-1, 93 — 168.

R. D. Van Valin, Jr. 1990. Semantic parameters of split intransitivity. *Language*, 66-2, 221 — 260.

H. Vogt. 1971. *Grammaire de la langue géorgienne*. Oslo: Universitetsforlaget.

Large Scale Analysis of Verb Subcategorization differences between Child Directed Speech and Adult Speech

Paula Buttery and Anna Korhonen
Natural Language and Information Processing Group
Computer Laboratory, Cambridge University
15 JJ Thomson Avenue, Cambridge, CB3 0FD, UK
{*paula.buttery*}/{*anna.korhonen*}@*cl.cam.ac.uk*

Abstract

Empirical data regarding the syntactic complexity of child directed speech (CDS) is necessary for determining its rôle in language acquisition. Of particular importance is data related to the predicate-argument structures and verb subcategorization frames (SCFs). However, manual analysis of SCFs is costly and consequently available data for evaluating theories is sparse. We address this problem by using the most comprehensive subcategorization system available to automatically acquire large scale empirical data related to verb SCFs from CDS (an edited corpus of the CHILDES database (MacWhinney, 1995)). We compare this data against adult speech (a subset of the spoken part of the British National Corpus (BNC) (Leech, 1992)) and find that SCFs typical to CDS are different and often simpler than those typical to speech between adults. We discuss the impact of our findings on the prevailing theories of language acquisition.

1 Introduction

Understanding the rôle, if any, of child directed speech (CDS) is of fundamental importance to language acquisition. Several manual small scale studies (see Snow (1986) for an overview) have suggested that CDS is very different from speech between adults: intonation is often exaggerated, a specific vocabulary can be used, and sometimes even specific syntactic structures. However, the rôle of CDS is by no means clear. Pine (1994), amongst others, speculates that the purpose of CDS is to merely engage the child in conversation. Snow (1986), on the other hand, suggests that CDS is actually teaching the child language. Clearly, larger-scale studies into the nature of CDS are required before we can begin to establish its rôle in acquisition. This paper details a systematic, large-scale investigation into the syntactic properties of verbs in CDS.

There is considerable evidence that syntactic information, in particular, is informative during language acquisition (e.g. (Lenneberg, 1967), (Naigles, 1990) and (Fisher *et al.*, 1994)). Often theories rely on syntactic diversity in the child's input for successful acquisition. For example, Landau and Gleitman (1985) suggest that children use verb subcategorization frames (SCFs) to identify novel word meanings; arguing that in many cases surface-structure/situation pairs are insufficient or even misleading about a verb's interpretation . Consider the sentences *Did you eat your cookie?* and *Do you want your cookie?* According to Landau and Gleitman the SCFs of *eat* and *want* cue their interpretations, i.e. *want* occurs with sentential complements, suggesting a mental component to its interpretation. Furthermore, they suggest that SCFs provide convergent evidence on the meaning of a verb. For instance, if *John zirks bill the book* the learner assumes *zirk* to be an active verb of transfer (such as *bring, throw, explain*), whereas if *John is zirking that the book is dull* the learner interprets *zirk* to be a mental verb.

Such a syntactically intensive theory of acquisition can only be supported if the input to children is sufficiently complex and diverse in its SCFs. In general, CDS is thought to be syntactically simpler than adult speech, using simpler and fewer SCFs (Snow, 1986). If the rôle of CDS is to teach language, as Snow suggests, then we may have a conflict with acquisition theories that require syntactic complexity and diversity.

Manual analysis of SCFs is very costly and therefore not ideal for large scale studies in specific domains, such as CDS. Automatic acquisition of SCFs from corpora now produces fairly accurate lexical data useful for (psycho)linguistic research (e.g. Roland et al. (2000)). However, these methods are yet to be applied to CDS.

In this paper, we address the problem by using the most comprehensive subcategorization system available for English to automatically acquire large scale empirical data related to verb SCFs from CDS. We use both qualitative and quantitative methods to compare the resulting data against that obtained from a corpus of adult speech. We discuss our findings in relation to the prevailing theories of language acquisition.

Section 2 describes our method for subcategorization acquisition and section 3 introduces the corpora we used in our work. Our experiments and results are reported in section 4 and section 5 provides discussion and summarises our observations.

2 Methodology

We used for subcategorization acquisition the latest version of Briscoe and Carroll's (1997) system (Korhonen, 2002) which incorporates 163 SCF distinctions, a superset of those found in the ANLT (Boguraev *et al.*, 1987) and COMLEX (Grishman *et al.*, 1994) dictionaries. The SCFs abstract over specific lexically governed particles and prepositions and specific predicate selectional preferences but include some derived semi-predictable bounded dependency constructions, such as particle and dative movement.

The system first extracts sentences containing specific predicates from a corpus. The resulting data is tagged, lemmatized and parsed using the 'RASP' system (Robust Accurate Statistical Parser; (Briscoe and Carroll, 2002)). Local syntactic frames including the syntactic categories and head lemmas of constituents are then extracted from parses. The resulting patterns are classified to SCFs on the basis of the feature values of syntactic categories and the head lemmas in each pattern. Finally a lexical entry is constructed for each verb and SCF combination whose relative frequency is higher than an empirically defined threshold.

3 Corpora

In order to make valid comparisons between SCF frequencies in CDS against adult speech there is a necessity to first ensure that the corpora are controlled for all other variables. Roland and Jurafsky (1998) have shown that there are subcategorization differences between written and spoken corpora, and furthermore that subcategorization is affected by genre and discourse type. Hence, we use only spoken data for both corpora and restrict data to conversation between family members and friends.

To ensure sufficient data for subcategorization acquisition, we have had to use an American English source for the CDS corpus although we had a British English source for the adult speech corpus. However, we do not expect this to be a problem: Roland *et al* (2000) have shown that subcategorization probabilities are fairly stable across American vs. British English corpora; finding any exceptions to be the result of subtle shifts in verb sense due to genre.

The following sections describe the two corpora we chose to experiment with.

3.1 Child Directed Speech - CHILDES Corpus

The CDS corpus has been created from several sections of the CHILDES database (MacWhinney, 1995): Demetras1 (Demetras, 1989b); Demetras2 (Demetras, 1989a); Higginson (Higginson, 1985); Post (Post, 1992); Sachs (Sachs, 1983); Suppes (Suppes, 1974); Warren-Leubecker (Warren-Leubecker, 1982). These sections of the database exhibit naturalistic interactions between a child and caretaker (average child age 2;7). Speakers are both male and female, from a variety of backgrounds and from several locations around the USA. Child speech has been removed from the corpus and there is no reading. The corpus contains 534,782 words and has an average utterance length of 4.8.

3.2 Adult Speech - BNC Corpus

Our adult speech corpus has been manually constructed from the demographic part of the spoken British National Corpus (BNC) (Leech, 1992) such that it contains friend/family interactions where no children were present. The speakers were recruited by the British Market Research Bureau and come from a variety of social backgrounds. Speakers are both male and female, from several locations around the UK and all have an age of at least 15. Conversations were recorded unobtrusively over two or three days, and details of each conversation were logged. The corpus contains 835,461 words and has an average utterance length of 7.3.

4 Analysis

4.1 SCF Lexicons

We took the two corpora and extracted from them up to a maximum of 5000 utterances per verb. To make the results comparable, an equal number of utterances per verb were used for both corpora. In practise this number was often determined by CHILDES, which was smaller of the two corpora. It was also affected by the highly zipfian nature of verb distributions (see e.g. Korhonen (2002)), i.e. the fact that most verb types are extremely infrequent in language.

4.2 Methods for Analysis

Both qualitative and quantitative methods were used to compare the data in two SCF lexicons. The similarity between SCF distributions in the lexicons was examined using various measures of distributional similarity. These include:

- Kullback-Leibler distance - a measure of the additional information needed to describe p using q, KL is always ≥ 0 and $= 0$ only when $p \equiv q$;
- Jenson-Shannon divergence - a measure which relies on the assumption that if p and q are similar, they are close to their average;
- Cross entropy - a measure of the information need to describe a true distribution p using a model distribution q, cross entropy is minimal when p and q are identical;
- Skew divergence - smooths q by mixing with p;
- Rank correlation - lies in the range $[-1; 1]$, with values near 0 denoting a low degree of association and values near -1 and 1 denoting strong association;
- Intersection - the intersection of non-zero probability SCFs in p and q;

where p and q are the distributions of SCFs in lexicons P and Q. For details of these measures see Korhonen and Krymolowski (2002).

In some of our experiments, the acquired SCFs were contrasted against a gold standard SCF lexicon created by merging the SCFs in the COMLEX and ANLT syntax dictionaries. We did this by calculating type precision (the percentage of SCF types that the system proposes which are correct), type recall (the percentage of SCF types in the gold standard that the system proposes) and F-measure:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (1)$$

4.3 Difference in Verb Types

Before conducting the SCF comparisons, we examined the 100 most frequent verbs in the BNC corpus versus the CHILDES corpus to get a more complete picture of the differences between the two data. We discovered that some verbs tend to be frequent in both corpora, e.g. *go, get, think, like, make, come, take*. However, closer analysis of the data revealed large differences. We discovered that in general, action verbs (e.g. *put, look, let, sit, eat, play*) are more frequent in CHILDES, while mental state verbs (e.g. *say, know, mean, suppose, ask, feel, seem*) - which tend to have richer argument structure - are more frequent in BNC. The 30 most frequent verbs in the two corpora are listed in Figure 1, in the order of their frequency, starting from the highest ranked one.

Rank	BNC	n	CHILDES	n
1	*get*	5000+	*go*	5000+
2	*go*	5000+	*be*	5000+
3	*say*	5000+	*do*	5000+
4	*be*	5000+	*see*	4200
5	*know*	5000+	*put*	4037
6	*do*	5000+	*get*	4018
7	*think*	4074	*want*	3411
8	*see*	2852	*can*	3409
9	*like*	2827	*let*	2771
10	*can*	2710	*look*	2585
11	*come*	2602	*think*	2280
12	*want*	2148	*like*	2038
13	*mean*	2078	*know*	1768
14	*look*	1930	*say*	1755
15	*put*	1776	*come*	1693
16	*take*	1443	*make*	1692
17	*tell*	1122	*okay*	1593
18	*make*	1092	*take*	1356
19	*use*	1016	*eat*	1172
20	*will*	1007	*give*	990
21	*give*	920	*play*	944
22	*buy*	590	*tell*	860
23	*leave*	548	*find*	661
24	*keep*	545	*happen*	581
25	*pay*	543	*sit*	580
26	*let*	536	*read*	571
27	*remember*	517	*remember*	563
28	*work*	495	*try*	556
29	*suppose*	489	*fall*	546
30	*play*	477	*will*	537

Figure 1: 30 most frequent verbs in adult speech (BNC) corpus vs. child direct speech (CHILDES) corpus

4.4 SCF Comparison

A subset of the constructed lexicons were compared for subcategorization similarities between the BNC corpus and CHILDES corpus. To obtain reliable results, we restricted our scope to 93 verbs—all those for which the total number of sentences analysed for SCFs was greater than 50 in both corpora, and which were thus less likely to be affected by sparse data problems during SCF acquisition. The SCF lexicons for these verbs were also contrasted against the gold standard described earlier in section 4.2.

The average number of SCFs taken by studied verbs in the two corpora proved quite similar, although verbs in BNC took on average a larger number of SCFs (19) than those in CHILDES (15).

	BNC	CHILDES
Precision	51.41	52.21
Recall	28.57	24.36
F Measure	36.73	33.22

Figure 2: Precision, recall and F Measure of CHILDES lexicon and BNC lexicon with respect to COMLEX-ANLT combined gold standard.

However, we found that most verbs (regardless of their frequency in the corpora) showed substantially richer subcategorization behaviour in the BNC than in CHILDES. A total of 80 frame types were hypothesised for the 93 studied verbs in the BNC, while 68 were hypothesised in CHILDES. The intersection between these frames in the corpora was not large (0.61).

To establish whether this difference was due to one lexicon being considerably less accurate than the other, we compared the SCFs in both lexicons against the gold standard. The results listed in Figure 2 show that the BNC lexicon had a slightly higher F measure than CHILDES: 36.7 vs. 33.2.[1] This was only due to the better recall of BNC (+4.21% compared with CHILDES), as CHILDES had a better precision than BNC (+0.80%). The differences in precision and recall – although fairly small – can be largely explained by the nature of SCFs in the two corpora. The smaller number of frames proposed in CHILDES were less complex and thus easier for the system to detect correctly, while the more varied SCFs in the BNC were more complex and also more challenging for the system.

Indeed the distributions of SCFs in the two corpora appeared fairly different. As shown in Figure 3, there was only a weak rank correlation between the frames in the distributions (0.46). The Kullback-Leibler distance denotes a low degree of correlation (1.0) and the results with other measures of distributional similarity are equally unimpressive (e.g. the cross entropy is 2.7).

	CHILDES vs. BNC
KL distance	1.022
JS divergence	0.083
cross entropy	2.698
skew divergence	0.533
rank correlation	0.463
intersection	0.608

Figure 3: Average similarity values

Our thorough qualitative analysis of SCF differences in the two corpora revealed reasons for these differences. The most basic SCFs (e.g. intransitive and simple NP and PP frames; which describe e.g. *he slept*, *he ate an apple* and *he put the book on the table*) appeared equally frequently in both corpora. However, a large number of more complex frames were either very low in frequency or altogether absent in CHILDES. For example, the verb *hear* appeared only in the following kind of constructions in CHILDES:

1. *I heard you*
2. *I heard*
3. *I heard that you came*

while in BNC it also appeared in the following kind of constructions:

1. *I heard it from him*
2. *Can you hear this out?*
3. *Did you hear whether he will come?*
4. *I heard him singing*

Several types of SCFs were poorly covered or largely absent in CHILDES. Many of these were frames involving sentential and predicative complementation (e.g. *I caught him stealing*, *he forgot what to do*, *I helped him to dress*) and verb-particle constructions (*I got him up from the bed*, *he came out poor*, *he looked it up*). Also a large number of adjectival frames were missing (e.g. *I remembered him as stupid*, *It dropped low*). On the other hand, frames involving prepositional or nominal complementation were covered fairly well in CHILDES (e.g. *I will get it from him*, *she built me this castle*).

While the SCF differences seem fairly big, they are not altogether arbitrary. Rather, they seem somewhat correlated with different verb senses and SCFs typically permitted by the senses. To gain a better understanding to this, we looked into Levin's taxonomy (Levin, 1993) which divides English verbs into different classes on the basis of their shared meaning components and similar syntactic (mostly subcategorization) behaviour. For example,

[1] Note that these figures are not impressive as performance figures, largely due to the fact that the gold standard was not fully accurate as it was obtained from dictionaries rather than from the corpus data. It was also too ambitious considering the size of the corpus data used in our experiments and the zipfian nature of the SCF distributions (i.e. many SCFs listed in large dictionaries were simply missing in the data, as the low recall indicates). However, the gold standard was adequate for the purpose of these experiments.

in Levin's resource, verbs such as *fly, move, walk, run* and *travel* belong to the same class as they not only share a similar meaning but also take similar SCFs.

When we compared for some of our test verbs the SCFs in the two corpora to those listed in Levin, we noticed that many of the SCFs absent in CHILDES and listed in the BNC and were just syntactically more complex manifestations of the same verb sense as that described by the CHILDES SCFs. For example, verb senses that take multiple sentential and predicative complements in Levin take just a smaller range of those SCFs in CHILDES than in BNC. However, some SCFs in BNC describe verb senses which were altogether absent in CHILDES. After a closer look, many of these senses proved to be extended senses of those exemplified in CHILDES.

In the light of this small scale investigation with Levin classes, it seems to us that to gain a better understanding of SCF differences in adult and CDS speech and the role of SCFs in language acquisition, it would be useful, in the future, to investigate to what extent SCF learning is mediated by the sense of the predicate and its membership in classes such as Levin's.

5 Observations

Some prevailing theories of language acquisition (e.g. that of Landau & Gleitman (1985)) suggest that verb SCFs provide convergent evidence on the meaning of a verb. These theories rely on the assumption that the frames provided in a child's input are adequately diverse to support learning. Meanwhile, Snow (1986) suggests that CDS plays an important rôle in the facilitation of acquisition. If Snow and Landau & Gleitman are both correct then we would perhaps hope to find that CDS is diverse in terms of its SCFs.

This appears to conflict with earlier small-scale empirical studies (e.g. (Snow, 1986)) which suggest that while CDS is quite complex (displaying, for example, the full range of conventional indirectness) it is syntactically much simpler than speech between adults. Our empirical results obtained from automatic SCF analysis of large-scale data[2] show conclusively that CDS is not only significantly simpler but also syntactically very different than speech between adults. Perhaps then, the rôle of CDS is to encourage the acquisition of simple frames, providing a basis from which more complex frames may be developed.

The fact that there is little correlation between the SCFs in two corpora is a little surprising as one might expect CDS to contain a subset of adult speech's SCFs. However, as our small scale experiment with Levin classes suggests, the SCFs seem nevertheless correlated via verb senses. While this issue requires further investigation, it is important to also note that some CHILDES SCFs absent in BNC may not be altogether absent in adult speech. Due to the Zipf-like nature of the SCF data, they may just occur in adult speech with a very low frequency. If this turns out to be the case after further larger scale experiments, it would indicate that most CDS SCFs are indeed a subset of those in adult speech but the frequencies of the SCF in the two corpora differ substantially.

Our results may also support Valian's (1990) findings that 4% of parental replies to children are ungrammatical, and 16% grammatical but not fully acceptable (examples from our CDS corpus include "play this together?", "another one missing."). Such utterances explain at least partly why there are SCFs present in the CHILDES lexicon that are missing from the BNC. Valian also found that adults tend to reply to children using an utterance which is lexically and structurally similar to the child's sentence (5% verbatim, 30% structurally similar). Since child speech at 2;7yrs (the average age of child subject in our CDS corpus) is usually simpler than adult speech ((Nice, 1925) and (Brown, 1973)) such repetition could help to boost the relative frequency of simpler frames in the CHILDES lexicon.

References

B Boguraev *et al.* 1987. The derivation of a grammatically-indexed lexicon from the longman dictionary of contemporary english. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 193–200.

E Briscoe and J Carroll. 1997. Automatic extraction of subcategorization from corpora. In *5th ACL Conference on Applied Natural Language Processing*, pages 356–363.

E Briscoe and J Carroll. 2002. Robust accurate statistical annotation of general text. In *Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499—1504, Las Palmas, Canary Islands.

R Brown. 1973. *A first Language: the early stages*. Harvard University Press, Cambridge, Mass.

M Demetras. 1989a. Changes in parents' conversational responses: a function of grammatical development. In *ASHA, St Louis*.

M Demetras. 1989b. Working parents' conversa-

[2] We will make our data publicly available via the web.

tional responses to their two-year-old sons. University of Arizona.

C Fisher *et al.* 1994. When it is better to receive than to give: syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92(1-4):333–375, April.

R Grishman *et al.* 1994. Comlex syntax: building a computational lexicon. In *International Conference on Computational Linguistics*, pages 268–272.

R Higginson. 1985. Fixing-assimilation in language acquisition. unpublished doctoral dissertation, Washington State University.

A Korhonen and Y Krymolowski. 2002. On the robustness of entropy-based similarity measures in evalution of subcategorization acquisition systems. In *Sixth Conference on Natural Language Learning*, pages 91–97, Taipei, Taiwan.

A Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge. Thesis published as Technical Report UCAM-CL-TR-530.

B Landau and L Gleitman. 1985. *Language and Experience: evidence from the blind child*. Harvard University Press, Cambridge, Mass.

G Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

E Lenneberg. 1967. *Biological Foundations of Language*. Wiley Press, New York.

B Levin. 1993. *English Verb Classses and Alternations*. Chicago University Press, Chicago.

B MacWhinney, 1995. *The CHILDES project: Tools for analysing talk*. Lawerance ErlBaum Associates, Hillsdale, NJ, second edition.

L Naigles. 1990. Children use syntax to learn verb meanings. *Journal of Child Language*, 17:357—374.

M Nice. 1925. Length of sentences as a criterion of child's progress in speech. *Journal of Educational Psychology*, 16:370—9.

J Pine. 1994. The language of primary caregivers. In C. Gallaway and B. Richards, editors, *Input interaction and language acquisition*, pages 13–37. Cambridge University Press, Cambridge.

K Post. 1992. The language learning environment of laterborns in a rural florida community. unpublished doctoral dissertation, Harvard University.

D Roland and D Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *COLING-ACL*, pages 1117–1121.

D Roland *et al.* 2000. Verb subcatecorization frequency differences between business-news and balanced corpora. In *ACL Workshop on Comparing Corpora*, pages 28–34.

J Sachs. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. In K Nelson, editor, *Children's Language*, volume 4. Lawrence Erlbaum Associates, Hillsdale, NJ.

C Snow. 1986. Conversations with children. In P Fletcher and M Garman, editors, *Language Acquisition*, pages 363–375. Cambridge University Press, New York, 2nd edition.

P Suppes. 1974. The semantics of children's language. *American Psychologist*, 29:103–114.

V Valian. 1990. Logical and psychological constraints on the acquisition of syntax. In L Frazier and J Villiers, editors, *Language Processing and Language Acquisition*, pages 119—145. Dordrecht, Kluwer.

A Warren-Leubecker. 1982. Sex differences in speech to children. unpublished doctoral dissertation, Georgia Institute of Technology.

Treating Support Verb Constructions in a Lexicon:

Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns

Silvie CINKOVÁ
Center for Computational Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
Praha 1, Czech Republic, 118 00
cinkova@ufal.mff.cuni.cz

Zdeněk ŽABOKRTSKÝ
Center for Computational Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
Praha 1, Czech Republic, 118 00
zabokrtsky@ufal.mff.cuni.cz

Abstract

We introduce a bilingual MR lexicon of Swedish support verb constructions that lemmatizes their noun components (predicate nouns). The lexicon is meant to be part of a valency lexicon of common Swedish verbs. It is based on the valency theory developed within the Functional Generative Description and it is enriched with Lexical Functions. In order to give the user some insight into event structure features of support verb constructions we concentrate on the morphosyntactic behavior of the predicate nouns and observe telicity in the entire constructions.

1 Introduction

This paper describes bilingual lexicographical processing of support verbs in a recently launched project of an XML-based Swedish-Czech lexicon of common Swedish verbs. The lexicon is meant to help advanced Czech learners of Swedish to master phrase-dependent uses of the commonest lexical verbs that often show a tendency to grammaticalization (as defined by (Hopper, 1987)) and further analyzed by (Heine, Claudi and Hünnemeyer, 1991), such as *sätta (put), ge (give), gå (go)* or *falla (fall)*. One of such grammaticalized uses of common verbs is their acting as support verbs. Support verb constructions are treated in a separate sublexicon, which is the issue of this paper.

2 Support Verb Constructions, Support Verbs, Predicate Nouns

Support verb constructions (SVCs) are combinations of a lexical verb and a noun containing a predication. From the semantic point of view, the noun seems to be part of a complex predicate rather than the object (or subject) of the verb, despite what the surface syntax suggests. Support verbs are understood as verbs occurring in SVCs. Predicate nouns are in general nominal parts of complex predicates (including SVCs).

3 Capturing SVCs in the Lexicon

3.1 Benefits of the SVC Lexicon for the users

An SVC is usually semantically transparent. Its meaning is concentrated in the noun phrase, while the semantic content of the verb is reduced or generalized. The matching verb is unpredictable, though often a metaphorical motivation can be traced back. Implicitly, SVCs affect the foreign language reception less than the production (Heid, 1998), (Malmgren, 2002) and (Schroten, 2002). Besides itemizing the commonest SVCs and giving their Czech translation equivalents, the lexicon aims at providing the users with relevant SVC-construction rules for varying communication needs with special regard to event structure.

3.2 Describing Verbs with a Lexicon of Nouns

If we look upon SVCs as collocations, the noun is the base, while the verb is the collocate - cf. e.g. (Malmgren, 2002), (Čermák, 2003) and (Schroten, 2002). Even in the cross-linguistic perspective it is the noun that constitutes the common denominator for equivalent support verb constructions, whereas the support verbs do not necessarily match.

Focusing on nouns both enables the enumeration of all verbs semantically related to the given noun together at one place and a more systematic description of restrictions in morphological number, article use and adjectival or pronominal modifications in the nouns. Inspired by (Hopper and Thompson, 1980), (Lindvall, 1998) and (Bjerre, 1999), we believe that morphosyntactic behavior of the noun together with lexical features of the support verb determine the event structure of the entire SVC in context.

3.3 Event Structure Hints in the Lexicon

SVCs are often referred to as one means of marking event structure in non-aspect languages. A kind of event structure opposition is assumed between a SVC and its corresponding synthetic predicate (when there is any). SVCs can emphasize inchoativity, durativity and terminativity. However, this gives no direct correspondence to the Slavic category of aspect which apparently is the product of more event structure features in combination, one of which being telicity.

(Lindvall, 1998) looks into transitivity, treating it as interplay between noun definiteness (not limited to article use) and verbal perfectivity, considering them two sides of the same coin. Telicity plays a substantial role in her inferences. Noun definiteness is a grammatical category in Swedish, as it employs articles in nouns, whereas verbal aspect is a grammatical category in Czech, as Czech (often) employs morphological means to express aspect in verbs. Swedish does not have the grammatical category of verbal aspect and Czech does not have the grammatical category of noun definiteness. Telicity is only a lexical semantic feature of verbs and verbal constructions in both languages.

In this lexicon, we try to gather relevant information about the interplay between a given Swedish support verb and its predicate noun to give the Czech user an idea about the structure of the event described by the entire SVC. We apply Hopper and Thompson's conception of transitivity, i.e. we do not confine the description to predicate nouns as direct objects of support verbs.

3.4 Telicity Marking of SVCs

Telicity, introducing values "telic" and "atelic" should be regarded as independent of "aspect" with its values "perfective" and "imperfective". More to this issue see (Nakhimovsky, 1996), e.g. p. 170n: "A verb lexeme is telic if a simple declarative sentence in the past tense in which that lexeme is the main predicate is a telic sentence. A sentence is telic if it describes a telic process. A process is telic if it has a built-in terminal point that is reached in the normal course of events and beyond which the process cannot continue." However, this definition does not require that the sentence must express that the terminal point has been reached. Whether the terminal point was reached or not is the information provided by the category of aspect which is independent of telicity. We mark the entire SVCs by "telic"/"atelic". An SVC is marked as atelic when both the event described by the predicate noun and the event described by the support verb are atelic, such as *ha besvär* (*have problems*). An SVC is marked as telic when

a) both the event described by the predicate noun and the event described by the support verb are telic, e.g. *fatta beslut (take a decision)*

b) the event described by the support verb is atelic and the event described by the predicate noun is telic, e.g. *dra en slutsats (draw a conclusion)*

c) the event described by the support verb is telic and the event described by the predicate noun is atelic, e.g. *få besvär (get problems).*

The event a) describes the termination of a process, and so does the event b) while the event c) describes the onset of a state, thus is inchoative (inceptive). When considering the SVC as a compound of a "verbal" and a "nominal" event it is obvious that the "nominal" event does not switch with the "verbal" event, as the "verbal" event does not actually "take place" due to the semantic depletion in support verbs (cf. (Fillmore, Johnson and Petruck, 2003)). It rather inherits some of the verb's semantic components, as shown by the examples. The examples also suggest a semantic opposition between an underspecified original state and the new state described by the "nominal" event. This implies that such SVCs are transitions. (Bjerre, 1998) claims: "SVCs denoting transitions are invariably achievements, either inchoatives or causatives; the SV always denotes an underspecified subevent$_1$." Transitions are telic events (cf. e.g. (Pustejovsky, 1991)). Hence SVCs like a), b) and c) will be marked as telic.

Still sticking to Nakhimovsky's definition of telicity, it is to be specified which of the two subevents in a compound event like transition is expected to be the telic process that has the "built-in terminal point that is reached in the normal course of events and beyond which the process cannot continue" (see above). To keep the premise that transitions are telic, the telic event must be the subevent$_1$ represented by the given support verb. The telicity marking refers to the entire SVC, but at the same time also to the support verb, no matter what the telicity conditions of its core meaning (i.e. of the most cognitively salient one) are like. E.g. *draw* would be intuitively classified as atelic when standing outside the context, but *draw* as a support verb in *draw a conclusion* would be telic.

4 Lexicon Architecture

The structure of the noun lexicon was mainly inspired by VALLEX, the FGD-based valency lexicon of Czech verbs (Straňáková-Lopatková et. al., 2002). Some features were taken from PDT-VALLEX, see (Hajič et al., 2003), which is a supporting lexicon for the manual annotation of the Prague Dependency Treebank. Though based on a lexicon primarily designed for treebanking, the lexicon of predicate nouns has neither been created on the basis of a Swedish treebank, nor is it intended for treebank annotation in the immediate future. The reason for choosing the (PDT-) VALLEX-like shape is that the formalized, yet still human-readable structure of (PDT-)VALLEX seems promising for keeping the desired level of consistency even when treating opaque linguistic phenomena. Applied to a FGD-based Swedish treebank, it would hopefully make a treebanking reference of the same quality as (PDT-)VALLEX. Unlike (PDT-)VALLEX, this lexicon is bilingual and sorts the collocational patterns of the nouns by Lexical Functions (Wanner, 1996). These additional features make its structure more complex.

4.1 Theoretical Background

4.1.1 Functional Generative Description (FGD)

The valency of the nouns has been described within the FGD framework. FGD is a formal stratificational language description framework, which makes use of achievements of the classical linguistics, going back to the functional-structural Prague School. For the purpose of this lexicon, we will concentrate on its tectogrammatical level that describes the underlying structure of a sentence, retaining the vagueness or indistinctness of the natural language.

Cross-linguistically, the – still language-specific – tectogrammatical representations of parallel texts are more similar than their surface syntax representations (which is supposed to be of benefit in machine translation (Hajič, 2002)). For more detailed description see (Sgall, Hajičová and Panevová, 1986) and (Panevová, 1980). The theory of FGD has been implemented in the Prague Dependency Treebank project (Sgall, Panevová, Hajičová, 2004).

In treebank annotation at the tectogrammatical level, only autosemantic lexical units are represented by nodes labelled by functors. A functor describes the semantic relation of each given node to its governing node. The left-to-right order of the nodes corresponds to the scale of communicative dynamism, to mirror the topic-focus configuration.

4.1.2 Lexical Functions (LF)

Lexical Functions are part of the Meaning-Text-Theory developed by Igor Mel'čuk and his collaborators (Mel'čuk, 1988), (Kahane, 2003). They enable a systematic description of "institutionalized" language-specific lexical relations in lexica for both human and computational use. There are two elementary types of LFs – paradigmatic and syntagmatic – and this paper concerns only the latter, which capture asymmetrical lexical relations. In terms of collocations, when two lexical units are collocates, one is usually the base that "selects" the other lexical unit to render a certain meaning together. The MTT captures it by the mathematical functional notation: LF_i (X) = Y, where X is called the keyword (the collocational base) and Y the value of the LF_i (the collocate). LFs can assign one value or a set of values to a given keyword. The values stand in the same lexical relation towards the keyword but they are not necessarily synonymous. The LFs describe the semantic relation between the keyword and the values. For examples and more details see (Wanner, 1996).

4.2 Entry Structure

This section gives a simplified description of the main elements and attributes of the lexicon microstructure as they are defined in the DTD. The elements are ordered as follows:

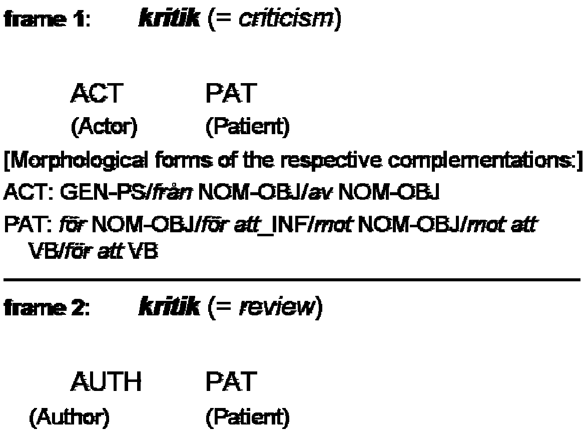
On the topmost level, the lexicon is divided into word entries. Each word entry relates to one headword lemma and its possible spelling variants. Homonyms get each an indexed word entry.

The element "Word Entry" comprises the elements "Headword lemma" and "Frame entry" (see Fig. 1). The former gives the lemma of the noun in question and its possible spelling variants. The latter describes the valency of the given reading of the lemma. Mostly each frame entry corresponds to one of the lemma's readings but when two semantically totally different readings happen to have identical frames, they are divided into two frame entries. This happens e.g. when the Czech translation equivalents differ in such an extent that they hardly ever can replace one another in the context.

4.2.1 Frame Entry

A valency frame is modelled as a sequence of frame slots. Each frame slot corresponds to one complementation of the noun in question. Each slot is assigned a functor according to its semantic relation towards the governing noun. Each slot includes an enumeration of its surface forms.

Surface forms of complementations of the given predicate noun are defined by the basic morphosyntactic categories, e.g. part of speech, gender, number, case, degree, definiteness and verb form. When the complementations are attached to the predicate noun by a preposition, the preposition is also recorded. The morphosyntactic categories are expressed by means of the SUC tagset (Ejerhed et al., 1992).



frame 1: ***kritik*** (= *criticism*)

ACT (Actor) PAT (Patient)

[Morphological forms of the respective complementations:]

ACT: GEN-PS/*från* NOM-OBJ/*av* NOM-OBJ

PAT: *för* NOM-OBJ/*för att*_INF/*mot* NOM-OBJ/*mot att* VB/*för att* VB

frame 2: ***kritik*** (= *review*)

AUTH (Author) PAT (Patient)

Fig.1: A word entry for *kritik* (*criticism*) with two valency frames. The first frame includes the surface forms of the complementations by means of the SUC tagset.

4.2.2 SVC-Frames

A sequence of SVC-frames nested in each frame-entry enumerates support verbs that typically occur with the given reading of the predicate noun in question (see Fig. 2). Each SVC-frame is defined by a combination of LFs (the basic and the complementary LFs, for more details see below) and by telicity conditions (attribute values "telic"/"atelic"). Each SVC-frame can comprise deliberately many support verbs. The verbs are displayed together with the predicate nouns as the entire SVCs, followed by Czech equivalents. Usually the sequence of SVCs with their Czech equivalents is followed by a sequence of example sentences taken from PAROLE, an ms-tagged Swedish corpus (http://spraakbanken.gu.se).

4.2.3 Lexical Functions in SVC-Frames

The lexicon of predicate nouns regards the predicate nouns as keywords of the basic Lexical Functions Oper$_1$, Oper$_2$, Labor$_{1,2}$, Copul and Func. Their values are by definition verbs.

In Oper, the predicate noun is a direct object of a transitive support verb, e.g. *pay attention*) or a prepositional object of an intransitive support verb, e.g. *get in touch*.

In Labor, the predicate noun is a prepositional object of a transitive verb, e.g. *subject sb to an interrogation*.

In Copul, the noun (or the adjective) is part of the predicate, in which a lexical verb has acquired a copula-like meaning, e.g. *fall ill. (= start to be ill)*.

In Func, the predicate noun is the subject of the verb, e.g. *The accusation came from John*.

The numbers denote indexes of the complementations (participants) of the events described. No. 1 is the Actor, No. 2 is the Patient. When an LF is specified by 1, it means that the Actor of the verbal event is identical with the Actor of the event described by the noun. When an LF is specified by 2, it means that the Actor of the verbal event is identical with the Patient of the event described by the noun.

Another LFs can complement the basic LFs, such as Phasal LFs, Causative LFs, the LF Anti and the LF Prox. These are LFs used by this lexicon. The Anti-LF is mainly stated when the negation of the predicate noun is not allowed to negate the SVC and other means have to be used instead, such as the negation of the verb or using a support verb with the opposite meaning. The Anti-LF is not being stated consequently due to the lacking lexical evidence.

4.2.4 Typical Morphosyntactic Representations of the Noun

The lemma noun itself is represented in the SVC-frame as a slot. The idea behind is that SVC-frames from the lexicon of predicate nouns will be interlinked with the not yet existing lexicon of common verbs. In the lexicon of common verbs support verb uses will be represented as separate valency frames with a special functor for predicate nouns. In the lexicon of predicate nouns the verb is not yet presented as a node but it is just listed as a text string within the noun frame.

Like any other slot, also the predicate-noun contains a set of SUC tags describing its morphosyntactic behavior in the given SVC (e.g. restrictions in number).

Two more attributes are attached to the noun slot to specify whether the predicate noun can be modified by an adjective or a possessive pronoun and to state the morphosyntactic conditions regarding the noun definiteness in combination with an adjectival attribute. (As in all major Germanic languages, article use is no longer an issue when the noun is determined by a possessive pronoun.)

The adjectival attribute would obtain the functor RSTR (Restrictive Adjunct), while the possessive pronoun would obtain the functor APP (Appurtenance). Following configurations can occur in Swedish:

- no attribute can be inserted into the SVC (RSTR_impossible)

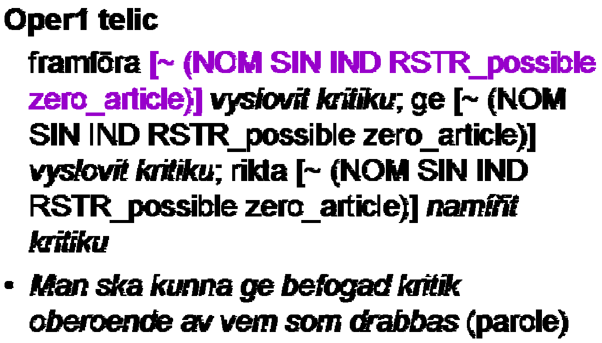
- an attribute can be inserted into the SVC and it can either be an adjective or a possessive pronoun (RSTR_possible)
- the attribute (adjective or a possessive pronoun) is obligatory (RSTR_obligatory)
- the obligatory attribute must only have the form of a possessive pronoun (APP_only_RSTR)
- no adjectival attribute can be inserted but the predicate noun can occur as part of a compound (compound_RSTR_only)

The conditions of indefinite article use in singular are rather complicated in Swedish SVCs, growing even more complex when an adjectival attribute is employed. Following configurations can occur:

- the predicate noun has no article without an attribute but gets it when employing an attribute (article_RSTR_dependent)
- the predicate noun never gets the indefinite article (zero_article)
- the predicate noun always has the indefinite article (article_obligatory)
- the predicate noun without an attribute can both occur with and without the indefinite article. (article_unrestricted).

Variations of the indefinite article in SVCs with an attribute combined with a restriction in the attributeless form were not considered, as they are unlikely to occur.

The morphosyntactic behavior of predicate nouns has been checked in the Swedish PAROLE corpus.



Oper1 telic

framfōra [~ (NOM SIN IND RSTR_possible zero_article)] ***vyslovit kritiku***; ge [~ (NOM SIN IND RSTR_possible zero_article)] ***vyslovit kritiku***; rikta [~ (NOM SIN IND RSTR_possible zero_article)] ***namířit kritiku***

- ***Man ska kunna ge befogad kritik oberoende av vem som drabbas*** (parole)

Fig. 2: The support verb entry nested in the first frame of *kritik*. It is defined by the Lexical Function Oper$_1$. It includes the telicity marking, the description of morphosyntactic characteristics of the predicate noun *kritik* in combination with the SVs *framföra*, *ge* and *rikta* (in square brackets after each verb), and Czech translation equivalents (in italics). Also an example with reference to the PAROLE-corpus is attached.

5 Further Work

So far, we have only worked with around twenty sample entries to refine the lexicon architecture and to make the editing more convenient. Thus we cannot present any statistical information, let alone any evaluation of the user's experinces at the moment.

The most frequent SVCs have been extracted from the PAROLE corpus and taken from three publications. The outstanding studies by (Dura, 1997), (Ekberg, 1993) and (Malmgren, 2002) together with our own investigations in the PAROLE corpus have already yielded a list of predicate nouns that should be included into the lexicon by manual lexicographical processing. We would also like to add frequency information into the lexicon. It was not attempted during this experimental stage, as quantitative analyses would be a rather tedious task in the unlemmatized PAROLE corpus. However, the authors kindly let us have the entire PAROLE. Thanks to that, we can try and get the corpus lemmatized before the lexicographical routine is seriously launched.

6 Conclusion

In Swedish there is apparently no way to infer aspect directly from the lexical features of the respective verbs. However, it is possible to state the telicity conditions and the common morphosyntactic representation of predicate nouns in SVCs, which could be a help to Czech speakers who easily get puzzled by the lexical way of expressing event structure in Germanic languages.

This lexicon description considers the interplay between the support verb and the predicate noun one of the more universal principles within the lexicon that (Pustejovsky, 2000) refers to as syntagmatic processes. We consider it possible to describe lexical features of the entire SVCs. The enumeration of possible article-adjective configurations (that are not synonymous precisely regarding the event structure) suggests possible event structure features of a given SVC when employed in context.

Even the very rough deconstruction and telicity notation of very few SVCs together with the notation of phasal and causative LFs suggests that SVCs can be a means of transforming states and processes into transitions. This already makes up a small hint for Czech speakers how to refine their way of expressing event structure by making their choices between a synthetic predicate and the matching SVC in Swedish.

Acknowledgements

This work has been kindly supported by GAUK 489-2004 and GAČR 405/04/0243.

References

T. Bjerre. 1999. *Event Structure and Support Verb Constructions.* In "Proceedings of the ESSLLI Student Session 1999.

F. Čermák. 2003. *Abstract Nouns Collocations: Their Nature in a Parallel English-Czech Corpus.* In "Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora". Birmingham.

E. Dura. 1997. *Substantiv och stödverb.* Meddelanden från Institutionen för Svenska Språket 18. Göteborg.

E. Ejerhed et al. 1992. The Linguistic Annotation System of the Stockholm-Umeå Corpus Project, Version 4.31. Publications from the Department of General Linguistics, University of Umeå, no. 32.

L. Ekberg. 1987. Gå till anfall *och* falla i sömn. *En strukturell och funktionell beskrivning av abstrakta övergångsfaser.* Lundastudier i nordisk språkvetenskap A 43.Lund.

Ch. J. Fillmore, Ch. R. Johnson and M. R. L. Petruck. 2003. Background to FrameNet. *FrameNet and Frame Semantics. International Journal of Lexicography* (Special Issue, Guest Editor: T. Fontenelle)16: 235-250.

J. Hajič. 2002: Tectogrammatical Representation: Towards a Minimal Transfer in Machine Translation. In "Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)", pages 216-226. Venice.

J. Hajič et al. 2003. PDT-VALLEX: *Creating a Large-coverage Valency Lexicon for Treebank Annotation.* In "Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Växjö, Sweden, November 14 - 15, 2003". pages 57-68.Växjö.

U. Heid. 1998. *Towards a corpus-based dictionary of German noun-verb Collocations.* In "Actes EURALEX'98 Proceedings". pages 301-312. Liège.

B. Heine, U. Claudi and F. Hünnemeyer. 2001. *Grammaticalization. A conceptual framework.* Chicago.

P. Hopper. 1987. Emergent Grammar. *BLS,* 13:139-157.

P. Hopper and S. A. Thompson. 1980. Transitivity in Grammar and Discourse. Language, 56:251-299.

S. Kahane. 2003. *The Meaning-Text Theory.* In "Dependency and Valency. An International Handbook on Contemporary Research". Berlin.

A. Lindvall. 1998. *Transitivity in Discourse. A Comparison of Greek, Polish and Swedish.* Lund.

S.-G. Malmgren. 2002. Begå *eller* ta självmord*? Om svenska kollokationer och deras förändringsbenägenhet 1800-2000.* Göteborg.

I. A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice.* New York.

A. Nakhimovsky. 1996. *A Case of Aspectual Polysemy, with Implications for Lexical Functions.* In: "Lexical Functions in Lexicography and Natural Language Processing". Studies in Language Companion Series (SLCS), Vol. 31. pages 169-179. Amsterdam-Philadelphia.

J. Panevová. 1980. *Formy a funkce ve stavbě české věty.* Praha.

J. Pustejovsky. 2000. *Syntagmatic Processes.* In "Handbook of Lexicology and Lexicography". de Gruyter.

J. Pustejovsky. 1991. The Syntax of Event Structure. *Cognition*, 41:47-81.

J. Schroten. 2002. *Light Verb Constructions in bilingual dictionaries.* In "From Lexicology to Lexicography". pages 83-94.Utrecht.

P. Sgall, J. Panevová and E. Hajičová. 2004. *Deep Syntactic Annotation: Tectogrammatical Representation and Beyond.* In "Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference", pages 32-38.

P. Sgall, E. Hajičová and J. Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht, Prague.

M. Straňáková-Lopatková et. al. 2002. Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In "LREC2002, Proceedings, vol. III. ELRA, pages 949-956.

L. Wanner. 1996. (Ed.) *Lexical Functions in Lexicography and Natural Language Processing. Studies in Language Companion Series (SLCS), Vol. 31.* Amsterdam-Philadelphia.

Defining a framework for the analysis of predicates

Montserrat CIVIT
University of Barcelona
mcivit@ub.edu

M.Antònia MARTÍ
University of Barcelona
amarti@ub.edu

Roser MORANTE
Tilburg University
r.morante@uvt.nl

Borja NAVARRO
University of Alicante
borja@dlsi.ua.es

Mariona TAULÉ
University of Barcelona
mtaule@ub.edu

Izaskun ALDEZABAL
University of the Basque Country
jibalroi@si.ehu.es

Abstract

In this position paper we present the research on verb predicates that we have carried out until now for Catalan, Spanish, and Basque, and we outline the framework of our future research, which is based on the idea that it is necessary to include syntagmatic and statistic information in lexical resources, such as WordNet, in order to use it in tasks of information extraction from annotated corpora, and in automatic syntactic and semantic tagging of corpora.

1 Introduction

The main goal of this position paper is to summarize the work on verb predicates that we developed in the last years from several perspectives: lexical semantics, corpus linguistics, and the semantic-syntax interface. Starting from that, and taking into consideration recent developments in the field, we sketch a working framework for the automatically semantic tagging of corpora taking advantage of the existing (but limited) semantic resources we have, and of the syntactic information they contain.

In section 2 we explain the motivations of our work, in section 3 we describe the work made until now and we evaluate it, and in section 4 we put forward a framework of future research.

2 Setting

This proposal is the result of several years of research in lexical semantics and corpus linguistics. In our analysis of Spanish, Catalan, and Basque verbs we found that studies about predicates in the fields of lexical semantics and the semantics-syntax interface show a lack of adjustment between the theoretical linguistic analysis and the real problems arising from automatic corpus processing.

Languages do not behave equally regarding the process of automatic analysis. The fact that English is a fixed constituent–order language allows a better adjustment between the theoretical description expressed in the computational lexicon and grammar, and the texts to be analyzed. Catalan, Spanish, and Basque are free constituent–order languages. This characteristic makes more complex the treatment of basic sentences and of the position of arguments. This is why the development of NLP tools, basically lexicons and grammars, has produced relatively poor results and has made evident the mismatch between theoretical work and real samples of language.

As it has been proposed in the literature, lexicons and grammars apart from containing linguistically motivated theoretical information should also incorporate information about coocurrence frequencies and about collocations. Our hypothesis is that adding this information makes the resources more efficient as NLP tools. Thus it is necessary to fill the gap between lexical resources and corpora by enriching lexical resources with syntagmatic information extracted from real samples of language.

3 Background

The authors of this article have worked in individual and in common projects related to NLP from different approaches: lexical semantics, analysis of predicates, and corpus analysis. In what follows a general description of these research directions is presented with the aim of justifying and establishing the basis for future work.

3.1 Lexical semantics

The knowledge sources and the training corpus for Basque, Catalan, and Spanish that were used in the Senseval–2 and Senseval–3 competitions have been developed adopting a lexical semantics perspective.

The quality of lexical resources used in the development of tagged corpus is one of the aspects that has been less taken into account in the Senseval competition. We carried out an experiment in which four different resources were evaluated: Minidir (Márquez et al., 2004), DRAE, EuroWordNet, and a dictionary based on the proposal by (Veronis, 2001). The aim of this experiment was to evaluate the quality of the resources and its effects in the results of the disambiguation tasks. The experiment consisted in letting the same corpus be annotated by three different annotators with each of the three dictionaries. The starting hypothesis was that the higher agreement between annotators would determine which was the lexical resource with more quality. As a result of this research it was found that the corpora with highest degree of agreement between annotators were the corpora annotated with the dictionary elaborated following (Veronis, 2001)'s model[1] and with MiniDir, a dictionary elaborated specifically for the Senseval competition, in which a criterium of minimum granularity of senses was applied. The average of senses per entry is 4. The degree of agreement in these cases was 90%. Both DRAE and EWN gave quite lesser results, between 60% and 70% of agreement.

In Senseval–3 the groups that had worked on Spanish showed a considerable improvement with respect to those of Senseval–2. Later it was found that the cause of the improvement was the methodology applied in the elaboration of the training corpus (Márquez et al., 2004): the same disambiguation system achieves better results if it is trained with the Senseval–3 corpus, than if it is trained with the Senseval–2 corpus.

In the context of Senseval–3, annotators were asked about the linguistic knowledge they were using when assigning senses to words. There was general agreement in considering that the syntagmatic information contained in MiniDir about collocations was essential, as well as the examples. Besides, all of them coincided in considering that the disambiguation of nouns and adjectives was resolved by looking at the strict local context, whereas for verbs it was necessary to identify subject and object.

From all this we deduce, firstly, that the syntagmatic information plays a main role in the disambiguation process, as already pointed out by (Veronis, 2001). Secondly, it seems necessary to have information about the subject and object in order to identify the sense of a verb. In (Nica et al., 2004) it is shown that the definition of syntactic patterns in a corpus and the extraction of paradigmatic information from them brings the corpus closer to the lexical resource and improves the quality of WSD systems.

3.2 Analysis of predicates

With the aim of elaborating a manual verbal classification based on syntactic–semantic criteria, we carried out several studies in the line of (Levin, 1993). The goal was to identify the diathesis in Catalan, Spanish, and Basque (Aldezabal, 2004), and to define verb semantic classes.

This work showed that establishing basic criteria to explain the relation between the diathesis and the verb senses is not straightforward, and that there are difficulties in distinguishing between diatheses and syntagmatic configuration. We produced a list of basic diatheses for 1.200 verbs of Catalan and Spanish, and for 100 verbs in Basque. In the case of Catalan and Spanish the 1200 verbs were grouped in two big classes: verbs of change (which accept the anticausative alternation) and verbs of transfer (which accept the underspecification of the trajectory component).

In the case of Basque, the same theoretical perspective was adopted, but instead of restricting the analysis to some verb classes, each verb was analyzed taking into consideration its occurrences in the corpus, as well as other diatheses alternations that Basque allows. This study showed that for syntactic alternations to have semantic classification power it is necessary to define in a declarative way what is an alternation, and the semantics it reflects (why certain structures form an alternation, which roles or semantic components do participate in it, which are the syntactic phenomena to take into account). Otherwise, when trying to identify the alternations for every verb in the corpus, doubts appear from the very first example.

In sum, those and recent studies for other languages (Schulte im Walde and Erk, 2005) showed that the problem of semantic verb classification, far from being solved, was becoming more complex, and that behind that problem lies the sense disambiguation problem.

3.3 Work with corpus

The authors have collaborated in the development of three treebanks with syntactic and semantic information (3LB corpus). The treebanks are three corpus of 100.000 words for Catalan, Spanish, and Basque, syntactically tag-

[1] Only four entries were elaborated following this model.

ged with phrases and functions. A subset of the corpus has been semantically tagged with WordNet (Palomar et al., 2004). In the Spanish 3LB corpus all nouns, verbs, and adjectives have been tagged. The Cast3LB corpus has approximately 1.400 verbs. In the Catalan corpus the same has been done for 10.000 words.

It is well known how difficult the elaboration of semantically tagged corpus is and how much human effort it costs. Although the data available are sparse, 3LB constitutes the first attempt at providing these languages with a corpus annotated with syntactic and semantic information.

The relation between the senses of a verb and the WordNet senses for subjects and objects can be automatically extracted from the 3LB corpus. Additionally, it is also possible to obtain the syntagmatic structures associated with every verb sense. Data sparseness is the problem that arises in this case, since in order to extract syntactic-semantic information the quantity of examples available is insufficient.

In addition to the 3LB corpus, the Spanish, Catalan, and Basque corpus developed for the lexical sample task in the Senseval–2 and Senseval–3 competitions are available. This corpus has 200 examples for each of the 50 words selected for the lexical sample task, which sums up a total of 10.000 sentences with only 1 tagged word. 10 of the 50 words are verbs (2.000 exemples).

4 Research lines

As it has been said above, we consider that for lexical resources to be useful in language analysis tasks (parsing), as well as in WSD task, it is necessary to enrich them with syntagmatic information. This is what the experiments carried out for tagging the corpus show. The problems that arise are how to acquire this knowledge automatically or semiautomatically, while at the same time guaranteeing its quality, how it will be coded later, and how it will be used.

It is our purpose to develop basic resources for Spanish, Catalan, and Basque in order to provide necessary tagged corpora that will allow carrying out machine learning experiments. In what follows, we propose some strategies based on automatic methods to create some of those resources.

4.1 Semantic disambiguation

Taking as a basis the material that we already have (projects 3LB and Senseval–3), our main goal is to perform an experimental study about the possible correlation between verb senses and semantic type of objects. The information obtained in this way will be added to shallow parsed corpora.

In order to do that we will start from the 3LB corpus (100.000 words), which has both syntactic (functions) and semantic information (synsets of WordNet) for the categories noun, verb, and adjective. For all the senses of verbs with a frequency rate higher than 20, we will extract the noun acting as head of the direct object (DO) and the associated synset. After that we will obtain its *specification mark* (Montoyo, 2002) for the list of direct objects of each sense. This is to say, we will find out in WordNet the lower synset (hypernym) that includes all or most of the synsets associated with the heads of the DO. Our hypothesis is that this node or specification mark defines, for every sense, a subset of EWN where candidates to DO can be found.

In order to verify the relevance of the results obtained, we will check in the Senseval-3 corpus (where only verbs are annotated with synsets) if there exist heads of NPs in the subset of EWN defined by the specification mark. If the result is positive, it will be considered a positive proof in the verification of the hypothesis. If the result is negative, the corpus Senseval–3 will be annotated syntactically, following the methodology defined for 3LB, with the aim of obtaining evidence about the correlation between the verb sense and the semantic type of the object. In the last case a wide collection of examples for every verb is available (a minimum of 200 examples), that might provide more evidence about the validity of the starting hypothesis.

If the results are positive, for the analyzed verbs it will be possible to use the resulting information with the following purposes: assigning the syntactic function *DO* to the NPs of shallow parsed corpora; assigning synsets to all DO on the basis of the specification mark; and semantically tagging the analyzed verbs.

We do not consider making the same study with subjects because in the three languages the subject is usually omitted, and because it is less determining in the semantics of verbs. A next step would be to analyze the prepositional arguments.

4.2 Enriching EWN with syntagmatic information

The information obtained from the previous experiments can be inserted in EWN. For every verbal synset it would be possible to express the nominal synsets that appear as an object, so that this information can be used in WSD processes.

4.3 Syntax–semantics interface

The 3LB corpus provides the necessary information to find out if it exists a correlation between syntactic structures and verbal senses. For example, for each main verb, the Cast3LB and Cat3LB corpora provide information like the specific sense of the verb, the main complements related to this verb, the kind of phrase, the syntactic function of the complements related to the verb, the head of each complement, and its specific sense. So, from these corpora it is possible to extract syntactic semantic patterns formed by each verb and their arguments (Navarro et al., 2004), and it is possible to develop a lexical data base of verb patterns.

Furthermore, we have designed a method for the interlingua alignment of patterns (based on the Interlingua Index of EuroWordNet), in which each pattern is related to the patterns of the same verb sense in other language. This method compares the semantic and syntactic features of each argument of each verb sense, and aligns them if there is syntactic and semantic consistency. With this approach, the diatheses problem is extended to a multilingual framework: indeed, one of the main problems in multilingual alignment of syntactic semantic patterns is that there are different diatheses alternations in different languages (Navarro et al., 2004).

This information can be helpful to complement the work already done about verb diatheses in Basque, Catalan, and Spanish. Starting from this basis it is possible to carry out a translinguistic study about how each language solves the expression of a diathetic expression.

5 Conclusions

In this position paper we have presented a methodology (4.1 and 4.2) for the syntactic and semantic tagging of corpora using information extracted from the 3LB multilingual treebank, and from the automatic analysis of predicates (4.3) of the three languages involved.

6 Acknowledgements

Part of the research presented in this paper has been financed by the following projects: Xtract2 (BFF2002–04226–C03–03) y 3LB (FIT–150500–2003–411).

References

I. Aldezabal. 2004. *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz.* PhD thesis, Basque Philology Department.University of the Basque Country, Leioa.

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

LL. Márquez, M. Taulé, M.A. Martí, N. Artigas, M. García, F. Real, and D. Ferres. 2004. Senseval–3: The spanish lexical sample task. In *Senseval–3 Third international Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 47–52. ACL, East Stroudsbourg PA–USA.

A. Montoyo. 2002. *Desambiguación léxica mediante marcas de especificidad.* PhD thesis, Univerity of Alacant, Alacant.

B. Navarro, M. Palomar, and P. Martínez-Barco. 2004. Automatic extraction of syntactic semantic patterns for multilingual resources. In *Proceedigns of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon.

I. Nica, M.A. Martí, A. Montoyo, and S. Vázquez. 2004. Combining ewn and sense-untagged corpus for wsd. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 5th International Conference, CICLing 2004 Seoul*, pages 188–200. Springer–Verlag, Heidelberg.

M. Palomar, M. Civit, A. Díaz, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.A. Martí, and B. Navarro. 2004. 3lb:construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. *Procesamiento del Lenguaje Natural*, 33:81–88.

S. Schulte im Walde and K. Erk. 2005. A comparison of german semantic verb classifications. In *Proceedings of the 6th International Workshop on Computational Semantics*, pages 343–353, Tilburg University.

J. Veronis. 2001. Sense tagging: does it make sense? In *Proceedings of the Corpus Linguistics Conference*, Lancaster, UK.

Unifying Lexical Resources

Dick CROUCH
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304
USA
crouch@parc.com

Tracy Holloway KING
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304
USA
thking@parc.com

Abstract

This paper describes our efforts to create a Unified Lexicon by extracting information from a variety of external resources, namely our XLE syntactic lexicon, WordNet, Cyc, and VerbNet. The UL is built in several steps: first the information is extracted from the resources; then it is merged into lexical entries based on word stem, syntactic subcategorization frame, meaning concept, and WordNet class; finally, patch files are run over the UL to create a cleaner version. The patched version of the UL is used to extract semantics to KR mapping rules, including default rules for where gaps occur in the external resources. This paper focuses on unifying lexical resources for verbs.

1 Introduction

There are a large number of external resources that have been developed to describe different aspects of the syntax, semantics, and abstract knowledge representation of verbs. Since these have been developed at different sites and for different purposes, they contain different types of information in different formats and cover different subsets of English. In order to exploit the information in these resources, it is necessary to merge the information and put it in a uniform format. This paper describes our efforts to build a Unified Lexicon (UL) with lexical entries for verbs based on their syntactic subcategorization in combination with their meaning as described by WordNet, Cyc, and VerbNet.[1]

For our purposes, the UL needs to be both machine and human readable. The machine-readability requirement comes from the fact that one of the main goals of these UL entries is to automatically extract rules which map from syntax to semantics to knowledge representation (Crouch, 2005). A second use of the UL is to determine where there are gaps in the resources and how best to create a series of defaults to fill these gaps. In order to do this, a linguist needs to be able to look through the UL for entries where there is missing information and then find similar entries that can be used to patch this information either with hand crafted rules or with defaults created from other information, usually information from other entries in the UL.

An sample entry in the UL is shown in fig. 1 for the HittingAnObject reading of transitive *hit*, as in *John hit the ball.* The UL entry contains information about WordNet class, Cyc knowledge representation, and VerbNet role mappings, role restrictions, and semantics. Each of these types of information forms a field in the entry, and the content of these fields can be extremely complex (e.g., the VerbNet field in fig. 1). There are also fields for comments, XLE lexicon information other than the subcategorization frame, derivational morphology information, and information from PARC internal resources. Not all fields need to contain information in a UL entry; part of the goal of building the UL is to see where gaps in information arise across the external resources.

The creation and use of the UL involves four steps: the data is extracted from the external resources; the extracted data is merged into the UL entries; the UL entries are corrected with hand-coded and automatically created patch files; mapping rules are extracted from the UL.

2 Extracting the Data

The current UL uses data from the XLE syntactic lexicon, a relatively complete research version of Cyc, VerbNet, and WordNet. We briefly describe these resources and some of the issues that arose when extracting the relevant data from them. In all cases, the data extraction

[1] Other external resources may be incorporated at a later date; these four resources (XLE lexicon, WordNet, Cyc, and VerbNet) were chosen because of their immediate relevance to the sem-kr mapping rules. ULs for other parts of speech are also planned.

```
(ul hit v v-subj-obj #$HittingAnObject
 (wnet ((wn 1172806 (verb contact)) (wn 1198410 (verb contact)) (wn 1359510 (verb contact))))
 (comments ())
 (xle ())
 (cyc (#$and (#$isa action #$HittingAnObject) (#$performedBy action subject)
  (#$objectActedOn action object)))
 (vnet ( (throw-17_1-1 Basic_Transitive
  ((role subj Agent ((int_control +))) (role obj Theme ((concrete +))))
  (sem ((motion (during E1) Theme) (exert_force (during E0) Agent Theme)
    (contact (end E0) Agent Theme) (not (contact (during E1) Agent Theme))
    (cause Agent E1) (meets E0 E1))))))
 (deriv ())
 (parc ()))
```

Figure 1: UL entry for HittingAnObject reading of transitive *hit*

is done automatically to allow us to easily update the UL when new versions of the external resources are released.

The XLE syntactic lexicon is a lexicon associate verb stems (∼9,700) with syntactic subcategorization frames (∼25,800 stem-frame pairs). It has been developed over the past several years as part of the broad-coverage English LFG grammar for the ParGram project (Butt et al., 2002). Extraction of the data simply comprised extracting each verb stem with its possible subcategorization frames. For example, from the entry in (1), we extract the information that *auction* can be either transitive (*They auctioned the goods*) or transitive with the particle *off* (*They auctioned the goods off/They auctioned off the goods*).

```
(1)   auction v
        { @(v-subj-obj %stem)
          @(subcat-source dict)
         |@(v-subj-obj_prt %stem off_)
          @(subcat-source byhand)}.
```

WordNet (Fellbaum, 1998) contains words, in our case only verbs, organized into synonym sets which represent underlying lexical concepts; these synonym sets are linked by relations such as hypernyms (e.g., *auction* is a type of *sell* which is a type of *exchange, change, interchange* which is a type of *transfer*). WordNet involved basically no direct extraction for the UL. However, WordNet class information is crucially used to determine whether entries from Cyc and VerbNet could be merged (section 3) and the information as to WordNet class(es) is recorded as being potentially useful in other aspects of the system, such as matching across representations. In addition, we anticipate that WordNet classes will play a crucial role in creating patch files (section 4) to fill in entries where there is not enough information from the other external resources to create useful sem-kr mapping rules.

2.1 Cyc

Cyc is a general knowledge base, including a large ontology of concepts and assertions about these concepts (Lenat, 1995).[2] Although Cyc contains information about concepts relating to many parts of speech, we initially extracted only the information known to be relevant to verbs. There were three main issues in extracting the Cyc data for inclusion in the UL.

The first concerns lemmatizing the verb forms. Cyc contains not just the base form of the verb, which is what is used in the UL entries, but also many inflected forms (e.g., in addition to listing an entry for transitive *push*, there will be duplicate entries for *pushes*, *pushing*, and *pushed*). To detect these duplicates, we put each verb form through the finite-state inflectional morphology that is used with the XLE English grammar. If this produced a stem with verbal tags that matched an existing verb entry from Cyc, then the form was discarded and only the lemmatized, base one was kept. As discussed in section 4, not all the verbs in Cyc were known to the morphology (e.g., *windsurfed*) and so some inflected entries had to be deleted with patch files.

The second issue involved the encoding of subcategorization frames in Cyc. These frames are labelled as to valency and sometimes phrase-structure type, but not usually with grammat-

[2] Cyc also contains a reasoning engine which is not used in construction of the UL.

ical functions. For example, the frame #$DitransitiveNPCompFrame indicates a verb which takes a subject and two additional NP arguments, such as *I gave him a book*. This must be mapped into grammatical functions as taking a subject, an object, and a secondary/thematic object (SUBJ-OBJ-OBJTH). In some cases, a Cyc frame might map into more than one grammatical function frame. Since there are relatively few frames listed per verb in Cyc, one of the purposes of the UL is to determine what strategies can be used to fill in Cyc-type KR for frames that are not listed. For example, if Cyc only listed the *that*-clause version of a verb and a *wh*-clause version was found in the XLE lexicon and/or VerbNet, could the *that*-clause information from Cyc be reasonably ported to the *wh*-clause one? Strategies for using the UL to fill gaps in external resources like Cyc are the subject of further research; however they must all make use of the patch file mechanism described in section 3.

A final issue with extracting the Cyc data involved the WordNet classes used in Cyc. Cyc associates the relevant WordNet class with a particular meaning of a verb. This information can be used to then associate these meanings with the relevant VerbNet meaning since VerbNet also include WordNet classes (section 2.2). However, two problems arose in doing this. The first was that Cyc uses an older version of WordNet than VerbNet. So, Cyc's WordNet class information had to be converted to the newer WordNet version. A second, more significant problem is that Cyc often uses the WordNet class for the relevant noun instead of verb. Given that Cyc is largely concerned with meaning and hence abstracts away from peculiarities of English syntax, this use of nominal classes for verbs is not unreasonable. However, the WordNet class numbers in these cases could not be used to merge the UL entries. Instead, these verbs had to be looked up in WordNet and then merged based on the retrieved information. The accuracy of the resulting merges is still being assessed, but initial inspection indicates accurate merges.

2.2 VerbNet

VerbNet (Kipper et al., 2000) classifies verbs according to Levin verb classes (Levin, 1993). It includes syntactic subcategorization information, information about thematic roles (e.g., agent, patient), and basic lexical semantics (see fig. 1). There were three main issues in extracting the VerbNet data for inclusion in the UL.

The first was converting VerbNet subcategorization frames into ones that were compatible with the XLE lexicon. This was difficult because the VerbNet subcategorization information is listed not as grammatical function information but rather as abstractions over the canonical phrase structure tree. For example, the frame corresponding to *present* in *I presented a solution to him*, is represented as in (2) (simplified from the original xml version).

(2) NP(Agent,[ ]), verb, NP(Theme,[ ]), Prep(to,[ ]), NP(Recipient,[ ])

From this, we extract the grammatical function specified subcategorization frame V-SUBJ-OBJ-OBL(to). To do this, we determine that the NP before `verb` is a subject, which will be linked to the Agent in the UL representation, and the NP immediately after `verb` is an object, which will be linked to the Theme. The NP following the `Prep` will be an oblique whose prepositional form must be *to* and this oblique will be the Recipient. This extraction becomes extremely involved for verbs which take NP small clauses, particles, expletives, or verbal complements.

The second issue in the VerbNet extraction was ensuring that a verb belonging to a particular VerbNet class inherited all the correct role restrictions from the classes above it. VerbNet classes frequently contain subclasses. Any role restrictions on the class also pertain to the subclass (sometimes nested several deep) and must be extracted accordingly. For example, the transfer_mesg-37.1 class which applies to sentences such as *Wanda taught French* has a restriction that its Agent is either animate or an organization. The subclass transfer_mesg-37.1-1 which applies to sentences such as *Wanda taught the students French* and its subclass transfer_mesg-37.1-1-1 for *Wanda taught the students* both inherit this restriction.

The final issue with VerbNet was that many verb frames have implicit roles. These roles are determined by looking at the semantics provided for the verb. If there is a thematic role mentioned that is preceded by a ?, e.g. ?Topic, then it is implicitly present in the verb frame and may have role restrictions on it. For example, the transcribe-25.4 class for *The secretary transcribed the speech* has an implicit Destination role which is restricted to being concrete. Note that this role is overt in other frames for this verb, as in *The secretary transcribed the speech into the record.*

To summarize, extracting the data from external resources into a format that we could then merge into UL entries involved a significant amount of work. Even for someone intimately familiar with all the resources, the conversion would have been non-trivial. Unfortunately, these resources are involved enough that an in-depth understanding of all of them is difficult and so much effort was spent on figuring out what should be extracted, converting it to a uniform format during the extraction, and then doing quality assurance on the results.

3 Merging External Data

Extracting data from a variety of sources and placing it in a moderately uniform format is unfortunately only part of the battle. The data from the different sources needs to be merged. The first stage of merging occurs in data extraction, by virtue of mapping XLE, VerbNet and Cyc verb entries to common subcategorization frames. However, both Cyc and VerbNet make what amount to sense distinctions for individual verbs within a particular sub-categorization frame. The principal task of merging these resources is therefore to identify equivalent Cyc and VerbNet sense distinctions. This is made harder in the case of VerbNet, since following the Levin verb classes it marks semantically significant syntactic alternations rather than alternative senses.

Both VerbNet and Cyc associate verb entries with WordNet sysnsets. These associations are used to help decide whether to merge Cyc and VerbNet entries for the same verb-subcat frame pairs. Unfortunately, this is not completely straight forward, for a number of reasons. (1) Cyc uses an older release of WordNet than VerbNet. (2) VerbNet uses only verb synsets, while Cyc often associates verbs with relevant nominal synsets. (3) Sense distinctions made by WordNet are often too fine for, and sometimes orthogonal to, ontological distinctions drawn between Cyc.

Part of the merging process attempts to recalculate synsets associated with Cyc entries, as a double check on the WordNet1.6 to 2.1 conversion process. This proceeds by identifying all words, of any part of speech, that map onto a particular Cyc concept, and collecting all the synsets for these words. This forms a very approximate cluster of synsets potentially associated with a Cyc concept, as opposed to the single synset allocated by Cyc. These clusters give a broader target when trying to match a Cyc entry up with a VerbNet entry. The alogorithm is greedy: if a VerbNet entry for a verb with a particular subcat frame has a synset that occurs in the Cyc cluster for the same verb-subcat frame pair, then it is assumed that the two entries should be merged. This sometimes results in multiple matches between a single Cyc entry and VerbNet entries, or vice versa. In such cases, multiple merged entries are produced.

Automated merging of verb senses is error prone. The patch file mechanism described in the next section provides a necessary means for correcting errors.

4 Patching UL Entries

In building the UL, we first extract the data from the relevant sources (XLE lexicon, WordNet, Cyc, and VerbNet) and then merge the information from these resources so that we have one entry for each stem, subcategorization frame, WordNet class, and meaning concept combination. Each such combination forms an id for that UL entry. This initial UL is then modified by patch files. These files can be produced by hand or automatically. The result is a new version of the UL and it is this version that the sem-kr mapping rules are extracted from.

Patch files are a convenient way of keeping a record of changes made to the UL after its initial extraction. It is important that the UL not be hand-edited directly. This is because the external resources from which the UL is constructed are themselves subject to change. We do not want to run the risk of losing hand-made modifications to the UL when rebuilding it to reflect a newer release of one of the external resources. By channeling all modifications to the UL through separate patch files, we can be sure to record any changes made

Patch files can be generated by automatically, semi-automatically or manually. But however they are generated, the format of a patch file is rigidly defined. A patch file consists of an ordered sequence of operations on UL entries, allowing them to be deleted, inserted, merged, or updated. Entries are identified by a key comprising (a) the word stem, (b) the part of speech, (c) the subcategorization frame, (d) the WordNet synset, and (e) a concept index derived from the Cyc knowledge representation of the word. In cases where some of the key information is missing (typically the concept index or the synset), null values are used.

Deletion is used to remove entries that are unwanted either because they are incorrect or because they will never be used in the mappings. For example, all of the inflected verb forms from Cyc that were not eliminated in the extraction (e.g., *snowbiking*) are deleted by a patch file. An example of an incorrect reading is that of the intransitive particle verb reading of *nod* for *He nodded off* which is incorrectly listed as #$NoddingOnesHead while it should only be listed with the meaning concept associated with falling asleep.

Insertion occurs when an entirely new entry is needed. Often, updating is used instead of insertion because existing underspecified UL entries, e.g. ones only with XLE lexicon and WordNet information such as *abbreviate* and *abdicate*, can be updated with the relevant additional information.

Merge merges two or more existing UL entries into a single new entry. Merges may be necessary where the WordNet classes did not align perfectly and yet the intended meanings of the two entries are identical. Often, an update to a UL entry will result in a new entry which can then be merged into an existing one. For example, if an incorrect subcategorization frame has been extracted from Cyc, this frame can be updated to the correct one and then the Cyc entry can be merged with an existing VerbNet one. This is done for many verbs taking prepositions since the encodings in Cyc and VerbNet were sometimes ambiguous between verbs taking obliques and those taking particles. In such cases, both were hypothesized in the original extraction and then updated and merged with a patch file based on the correct analysis.

Updating replaces a specified field in the UL with a new one.[3] Three operations are possible: adding, removing, and replacing. Each of these operates on a specified field in the UL. The fields include the word itself, the subcategorization frame, each of the types of extracted information (e.g., VerbNet), and a comment field. Adding creates a value for a field where there was none before. This can be used to insert comments into the UL entry. For example, many VerbNet entries have oblique arguments that the XLE grammar analyzes as adjuncts. For example, the XLE lexicon has a transitive use of *punch* but not one which takes an object and an *on* oblique (e.g., *He punch him on the arm*). For these verbs, a comment is inserted stating that the oblique is an adjunct and this comment allows the extracted sem-kr mapping rules to look for the appropriate adjunct grammatical function instead of an oblique one. Removing deletes the information in a given field. For example, if the VerbNet information for a given verb was incorrect but the Cyc and XLE information was correct, the UL entry could be updated by removing the VerbNet field. Finally, replacing removes the existing value for a field and replaces it with a new one. This is used to turn the Cyc multiword verbs into their single word equivalents. For example, the word *breathe in* is replaced by *breathe* and simultaneously its intransitive subcategorization frame is replaced by the intranstive frame with an *in* particle. This new entry can then be merged with the existing UL entry for that reading of the verb.

To summarize, a system of patch files is available to modify the UL from its initial state in which only information extracted from the external resources is used. Patch files can delete, insert, and merge entries, as well as modify any field in the entry. Since the rules in the patch file are ordered, entries are often modified and then merged to create single, accurate UL entries with information unified from all of the external resources.

5 Results and Conclusions

This paper describes our efforts to create a Unified Lexicon by extracting information from a variety of external resources, namely the XLE syntactic lexicon, WordNet, Cyc, and VerbNet. The UL is built in several steps: first the information is extracted from the resources; then it is merged into lexical entries based on verb stem, syntactic subcategorization frame, meaning concept, and WordNet class; finally, patch files are run over the UL to create a cleaner version. The patched version of the UL is then used to extract sem-kr mapping rules, including default rules for where gaps occur in the external resources.

The current UL contains 45,704 entries for 9,835 verb lemmata. 22,208 have no VerbNet information. 42,160 have no Cyc information. Of these, 22,122 have neither VerbNet nor Cyc

[3] There is also an update and copy command that copies the entry and then only updates the copy, leaving the original entry as well. This is often used to split entries and then merge them with several other entries.

information (e.g., *adapt*); that is, they effectively only contain the information from the XLE syntactic lexicon and WordNet. 17,991 have syntactic frames which came from VerbNet and were not in the XLE lexicon; the majority of these are frames with multiple oblique PP arguments (e.g., *The witch turned him from a prince into a frog*) and various types of resultatives (e.g., *Linda taped the box shut*) and middles (e.g., *Labels tape easily to that kind of cover*).

There is still much work to be done to fully exploit the UL in our syntax to semantics to KR mapping system. The next task is to extract mapping rules from the UL and incorporate them into the sem-kr mapping system. Then patch files need to be created to systematically fill in some of the gaps in the UL. Since there are many entries with VerbNet information but no Cyc information, we hope to use the VerbNet information to make informed guesses as to the Cyc meaning of the verb. In addition, WordNet classes may be used to determine the closest synonym for a given verb and the entry for that synonym could then be used to augment the UL entry for that verb.

Longer term work includes the incorporation of other external resources into the UL (e.g., derivational morphology, ComLex, FrameNet). In addition, ULs are being created for other parts of speech, including nouns and adjectives. The immediate need for these in our system is less pressing than for the verb UL described here because the external resources, in particular Cyc, can be used directly as a temporary measure.

6 Acknowledgements

This work was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

References

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar project. In *Proceedings of COLING2002, Workshop on Grammar Engineering and Evaluation.*

Dick Crouch. 2005. Packed rewriting for mapping semantics to KR. In *Proceedings of the 6th International Workshop on Computational Semantics.*

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* The MIT Press.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *AAAI-2000 17th National Conference on Artificial Intelligence.*

Doug Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. In *CACM 38, No. 11.*

Beth Levin. 1993. *English Verb Classes and Alternations.* Chicago University Press.

Predicting verbs in verb final clauses?

Dieuwke de Goede
University of Groningen
Department of Linguistics, PO Box 716
9700 AS Groningen, The Netherlands
d.de.goede@let.rug.nl

Roelien Bastiaanse
University of Groningen
Department of Linguistics, PO Box 716
9700 AS Groningen, The Netherlands
y.r.m.bastiaanse@let.rug.nl

Abstract

We present a Cross-Modal Lexical Priming experiment that shows priming of verbs before their actual occurrence at the end of main clauses in Dutch. We suggest that these results indicate that listeners make predictions about the upcoming verb when processing verb-final constructions. Post-hoc off-line tests fail to support the alternative hypothesis that the effect indicates direct priming from the object head noun (lexical priming).

1 Introduction

In certain circumstances people are able to predict (or: anticipate) upcoming sentential information quite successfully (e.g. in the 'cloze procedure', where participants have to complete a sentence fragment). For on-line processing, ERP experiments (starting with Kutas and Hillyard, 1984) have repeatedly demonstrated that a less plausible[1] or less expected upcoming word results in a greater amplitude of the N400. However, most ERP experiments showing this effect used a self-paced reading paradigm or presented spoken words with intervals between the words. It is a matter of debate whether prediction also plays a role when sentences are spoken rapidly and fluently, when the situation does not 'encourage' these processes.

In this paper we will review some literature suggesting that prediction indeed plays a role during more 'natural' language processing. We will try to expand this literature, which mainly addresses the prediction of nouns, to some recent findings on the prediction of verbs. Finally, we will present new data that suggest that in verb-final sentences, verbs can be predicted during ongoing spoken sentence processing.

We do not use *prediction* in the strict sense of the word. Rather, we assume that during sentence or discourse processing all information is used incrementally to restrict the plausible semantic domain of upcoming words (in line with e.g., McRae et al., submitted; Van Berkum et al., in press). The more information that has been processed and the more relevant this information has been, the more the semantic field of an upcoming word can be narrowed down.

We contrast prediction to *lexical priming*. Lexical priming takes place at the level of individual words: an incoming word is processed faster when it is associatively related to a previous word.

2 Predicting nouns

A verb imposes both syntactic and semantic constraints on its arguments, and it has been found repeatedly that during on-line sentence processing these constraints are accessed immediately upon encountering the verb. Verbs that have a more complex argument structure take longer to process (Shapiro et al., 1991) and arguments that occur later in the sentence are checked against these constraints (Friederici & Frisch, 2000). It is only a small additional step to assume that the processor uses the information released by the verb to set up predictions on information that has yet to come.

Indeed, evidence has been found for prediction of nouns on the basis of the preceding sentence context in a visual world experiment (Altmann & Kamide, 1999). In this paradigm participants listen to sentences and at the same time inspect a semi-realistic visual scene. The results suggested that information about the verb (sometimes in combination with the subject) is used to restrict the domain of plausible arguments (direct objects) that are to follow the verb.

Recently, Van Berkum et al. (2002, in press) showed that in certain circumstances people can even use discourse information to narrow down the semantic domain to *one specific* noun. In an ERP experiment they studied Dutch spoken two-sentence stories that ended in either an expected noun, or a perfectly possible, but less expected noun (based on off-line cloze tests). These nouns always differed in gender, which in Dutch influences the inflection of the adjective. When comparing the unexpected inflection condition with the expected inflection condition, results showed an early positive deflection emerging

[1] In most ERP research plausibility is operationalized as 'cloze probability': the proportion of participants that respond with the same word (in a cloze test)

directly after the inflection, *before* the onset of the noun, so *before* the standard N400 effect occurred at the noun itself.

To exclude the possibility that these effects were caused by lexical priming from words in the preceding discourse context to the relevant word, the original sentences were compared with sentences where the same potential prime words were used in such a way that the message that the discourse conveyed caused both final nouns to be equally *un*expected according to off-line cloze tests. In the latter sentences, no differential effects were found in ERP waves (Otten & Van Berkum, 2004).

The previous studies have in common that a noun is predicted at a point in time when the verb already has been processed. However, in many languages the verb only appears at the end of the clause or sentence. A relevant question is thus whether nouns can play a constraining role in sentence processing as well, and whether this could lead to prediction effects for verbs comparable to the ones found for nouns.

3 Predicting verbs

In the literature, only indirect or inconclusive evidence can be found for verb prediction in sentence context. Two studies tested priming for verbs in isolation. Salverda et al. (2004) showed that subjects and objects depicted in a visual scene can prime spoken verbs. This was the case even when the subjects in the scene were not actually engaged in the action that the verb described. The effects could be explained by lexical (object-verb) priming as well. Therefore, in a further study the potential subjects present in the scene were manipulated such that one event representation was more plausible than another one. This study suggested that effects found using the visual world paradigm are more likely to be caused by an interpretation of the combination of things present in the visual scene than by simple lexical object-verb priming.

McRae et al. (submitted) used a naming task to test whether a single noun could lead to the prediction of a certain class of events (verbs). They showed that expectancies can be generated from typical agents, patients, instruments, and locations, resulting in facilitation of naming times for verbs denoting the event.

At the sentence level, Hoeks et al. (2004) found that when a sentence-final verb fitted poorly with the preceding sentence context, the N400 effect was stronger in strong constraining sentences than in weak constraining ones even when the context contained exactly the same words, suggesting that the effect cannot be explained solely by lexical priming. However, these results were obtained in a word-by-word reading ERP experiment, which might encourage prediction.

Kamide (2004) found a pattern of results that can be tentatively interpreted as evidence that people predict semantic properties of the forthcoming verb in English object relative clause constructions. In a visual world paradigm she compared two constructions: *The cake which the boy will* *eat/move* *soon was made for his birthday.* The sentences were presented aurally while the participant was inspecting a visual scene with a boy, a cake, a ball, a toy car, and a toy train. In the *eat* condition, the cake was the only object in the scene that matched the verb's semantic information (being edible). In the *move* condition, the cake was not the only movable object. The study was designed to look at the effect of gap-filling: the direct object *cake* is the filler, which is presumably accessed at the gap, directly after the verb. However, the results showed that directly after the verb, when the word *soon* was processed, there were more looks to the cake in the *move* condition than in the *eat* condition. Kamide suggests that the preceding sentence context is used to predict certain properties of a possible verb. Immediately after the verb then, the prediction is evaluated against the incoming information. When the evidence does not go with the prediction, a 'surprise' effect occurs, manifesting itself as additional looks to the antecedent.

Although many relevant questions concerning predictive verb priming have been addressed in the studies discussed so far, they have failed to bring all relevant issues together. McRae et al. and Salverda et al. did not conduct their research at the sentence level. Hoeks et al. studied word-by-word reading. Kamide is the only one who used a paradigm with spoken language processing. However, she could only present indirect evidence for prediction effects.

In the current paper, we show that during *on-line spoken sentence processing* the combination of a relevant agent and patient (and a modal verb) can prime the clause-final verb. In contrast to McRae et al. (and probably also Salverda et al., and Hoeks et al.) we did not use (proto-) typical agents and patients. Our materials were originally designed in such a way that the different sentence parts were as unrelated to the main verb as possible, while allowing the whole sentence to remain as natural and plausible as possible.

4 The current experiment

We present a Dutch Cross-Modal Lexical Priming (CMLP) experiment, where participants listened to sentences and made a lexical decision to a visual probe presented at a particular point during each sentence. In CMLP facilitation of reaction times to a probe that is associatively related to a particular word in the sentence as compared to reaction times to a probe that is unrelated (but matched to the related probe) is attributed to priming effects. If priming is found at a certain point during the sentence, this is taken as evidence that the meaning of the relevant word in the sentence is activated. One advantage of the CMLP task is that, if implemented correctly (low proportion of related probes, enough variation in presentation point of probes, enough fillers, etc.), it reduces the likelihood that participants engage in prediction strategies. Also, if the sentences are presented at a normal speech rate, the probes are typically not integrated into the sentences.

4.1 Method

4.1.1 Participants

41 undergraduate and graduate students from the University of Groningen (all native speakers of Dutch) participated in the experiment.

4.1.2 Materials

Experimental sentences were of the following structure: subject NP – modal verb – object NP – adjunct (adverbial phrase of time) – main verb – conjunction – new clause (see 1).

(1) De kleine jongetjes zullen de fanatieke voetbaltrainer elke zaterdag[1]ochtend weer imiteren [2], want ze [3] willen later allemaal profvoetballer worden.
The little boys will the fanatical soccer coach every Saturday[1]morning again imitate [2], because they [3] want to later all pro soccer player become.

The visual probes were verbs that were either associatively related to the main verb (*nadoen = to copy*) or unrelated (*filmen = to film*), but matched as well as possible to the related probe as to pretested baseline lexical decision time, frequency, length and argument structure[2]. We used the same prime - related probe - unrelated probe triads as were used in De Goede et al. (submitted). Probes were presented at three different positions (see example sentence): probe point [1] was placed 700 ms after the onset of the adjunct (i.e. after both arguments have been read), probe point [2] at the offset of the main verb (the distance between probe point [1] and [2] was on average 1240 ms), and probe point [3] 700 ms after probe point [2], on average 153 ms after the offset of the conjunction.

There were 41 experimental sentences and 42 pseudo-experimental sentences (sentences with the same structure as the experimental sentences). The pseudo-experimental fillers were combined with non-words, to prevent any correlation between sentence type and response type (word/non-word). In addition, 20 filler sentences of different structures (10 words, 10 non-words) and 15 yes/no comprehension questions were added (to encourage participants to pay attention to the spoken sentences).

A completely counterbalanced design was created to assure that all participants saw both related and control probes, and saw probes at all three probe points. Each participant was tested twice, on the same list, but with related and control probes shifted. There were at least two weeks in between the two sessions.

4.1.3 Procedure

The participants were tested individually in a sound-proof room with no visual distractions. The sentences were presented over headphones with an inter stimulus interval of 1500 ms. The probes were presented on a standard computer screen. The experimental software Tempo (developed at the University of California, San Diego, for running CMLP-studies), combined with a response box with two buttons, was used to present the items and register the accuracy and RTs of the responses. Each probe was presented for 300 ms and a response could be given within a 2000 ms interval from stimulus onset. Importantly, the sentences continued without interruption during visual presentation of the probe.

Participants were instructed to listen carefully to the sentences and to expect comprehension questions after some sentences. Questions were answered and lexical decisions were made by pressing the left button on the button box for ‘no’ and ‘non-word’ and the right button for ‘yes’ and ‘word’. Participants were instructed to answer as quickly and accurately as possible.

4.2 Results and Discussion

Participants were excluded from further analysis if their error score on the lexical decision task was greater than 10%, if their mean or SD reaction times (RTs) deviated from the overall mean or SD by more than 2.5 SD, or if less than 67% of the comprehension questions were answered correctly. Data from four participants were excluded for these reasons.

[2] Unfortunately, for 5 items the control probe was intransitive, while the related probe was transitive. However, the results did not change when these 5 items were left out.

Error rates were low (2.0%) and equally distributed across related and control probes and across probe points. The exclusion of errors and outliers (> 2.5 SD) resulted in 3.4% percent data loss.

The mean RTs for all probe points and probe types are presented in Table 1 (the values that are presented are derived from the subject-analysis; the item-analysis revealed very similar data).

Probe position Probe type	pp [1]	pp [2]	pp [3]
control	701	702	699
related	685	678	690
difference	*16**	*24**	*9*

* $p < .01$ (paired samples t-test, subject analysis)

Table 1: Mean reaction times to related and control probes at each probe point.

Both subject- and item-based ANOVAs revealed a significant main effect of probe type (priming); overall, the related probes generated shorter RTs than the control probes: F_1 (1,40) = 14.55, p < .001; F_2 (1,40) = 6.87, p = .012. There was no significant interaction between probe point and probe type (F_1 (2,80) = 1.53, p > .2; F_2 (2,80) =1.05, p > .3). Planned comparisons showed no interaction effect between probe point and probe type for probe point [1] and [2] (F_1 and F_2 < 1). The interaction for probe point [2] and [3] did not reach significance either (F_1 (1,40) = 2.77, p = .104; F_2 (1,40) = 2.18, p = .148).

The results are in line with verb prediction: at probe point [1], after the processor has encountered the subject NP, modal verb, object NP and part of the adjunct, but well before the occurrence of the main verb, significant priming is found for related as compared to control probes: t_1 (40) = 3.27, p = .001; t_2 (40) = 1.79, p = .041.[3] The results at probe point [2] replicate earlier findings where priming of related versus control probes was found directly after the verb. In the current experiment there was a 24 ms advantage for the related probes (t_1 (40) = 2.99, p = .003; t_2 (40) = 2.56, p = .007). The effect at probe point [3] also replicates earlier findings, where the activation of the verb always dissipated in the embedded clause, although the decrease in priming is not as strong as it was in earlier experiments: t_1 (40) = 1.34, p = .10; t_2 (40) = 1.27, p = .11. Possibly, there is a small spill-over effect from the main verb itself, as we measured only 700 ms after its appearance in the current experiment.

4.3 Discussion of the pre-verbal effect

Although the priming effect at the pre-verb probe point can be interpreted to indicate anticipation of the verb or a class of concepts with which the main verb overlaps enough to find priming, other explanations are in line with the data as well. The most obvious possibility is that what we found is a simple lexical priming effect from the object head noun to the main verb, causing faster RTs to probes related to this verb than to control probes, even though we attempted to select plausible arguments with no associative relationship to the verb.

In earlier CMLP experiments it has been shown that nouns (mainly direct objects) deactivate quickly in a sentence context (according to Featherston (2001) the existing data (mainly on English) converge on a figure of about 500 ms). Although in our experiment the distance between object HN and probe was more than 500 ms (700 ms), it is worthwhile to exclude this possibility and show that, on the contrary, prediction is a better explanation for the results.

In the following section we will therefore try to show that the effects are suggestive of prediction of the verb and not of simple lexical priming by the object head noun. As this experiment was not set-up explicitly to answer this question, we will do this by means of post-hoc tests. We will present the data of a cloze test and a relatedness test.

4.4 Post-hoc off-line tests

4.4.1 Cloze test

In a paper-and-pencil test 37 participants were presented with fragments of the experimental sentences. The fragment that is relevant here consisted of the main clause up to the object HN, so without the adjunct and the main verb. Participants had to write down a short ending (1-5 words) for each fragment. The materials were counterbalanced across four lists (10, 10, 9 and 8 participants per list), such that each participant saw only one fragment of each experimental sentence, and saw all four fragment types. 68 Fillers were included in each list.

The answers of the participants were rated on a simple 2-point scale by the first author (the ratings were checked by an independent observer), where 2 indicated that the response contained either the exact final verb used or the exact related probe, a 1 indicated that the verb in the response was closely related to the final verb (and thus to the related probe), and a 0 indicated that the verb in the answer was not clearly related to the prime and

[3] As no inhibition effects were expected all t-tests are 1-tailed.

related probe. For each sentence the mean scores per fragment type were calculated.

4.4.2 Relatedness test

In a paper-and-pencil test 10 participants were asked to rate for each sentence the degree of relatedness of the object HN with the related probe and with the control probe. The underlying assumption is that lexical priming occurs if two words are related.[4] Relatedness was rated on a 7-point scale, with 1 being very unrelated and 7 very related. 16 Fillers were added for which either one of the two or both probes were clearly related.

4.4.3 Results and discussion

The mean cloze value was .47 on a scale ranging from 0 to 2 (minimum score: 0, maximum score: 2). For 25 items the mean cloze value was lower than the overall mean of .47. For these items, the priming effect at the preverbal probe point was 6 ms (n.s., t (14) < 1), as opposed to 16 for all items together. For the 15 items that had mean cloze values above the overall mean, the on-line priming effect was 34, which was significant: t (14) = 2.63, p = .01.[5]

We than checked whether there was a significant difference in relatedness score for the two item groups that were identified on the basis of the cloze test. This was not the case, the mean relatedness ratings for both the related and the control probes were very similar (mean scores for related probes: resp. 4.1 and 4.2, and mean score for control probes: resp. 2.6 and 2.5; both t's (38) < 1).

These results suggest that the items for which participants agree off-line on the semantic class of the verb that is to be expected at the end of the clause (that is, they come up with either the related probe, the prime verb, or a clearly related other verb), are exactly those items that show pre-verbal priming effects. And importantly, the difference in priming effect between the high- and low-cloze value groups cannot be attributed to any difference in off-line relatedness scores.

5 General Discussion

In a Cross-Modal Priming Experiment we presented Dutch spoken sentences in which the main verb appeared at the end of the main clause. We found facilitation in reaction times to visual probes related to the main verb (as compared to unrelated matched controls) *before* the verb was actually encountered. We suggest that this finding can be best explained as a predictive priming effect. On the basis of all preceding sentential information (the subject NP, a modal verb, and the object NP) the parser is in some cases able to impose constraints on the semantic aspects that the main verb should have, resulting in priming for verbs whose semantic characteristics match the activated class of concepts.

Anticipation of verbs has been shown before on a more structural, syntactic level, in studies where the number of arguments preceding the verb is manipulated and the effects on verb integration are measured (e.g., Ahrens, 2003; Konieczny, 2000). Our data, however, are interpreted in line with recent work showing prediction effects at a semantic or conceptual level.[6] We show that when the preceding information in the sentence can be used *off-line* to make adequate predictions about the semantic aspects of the verb that will occur at the end of the clause (cloze test), the parser uses this information *on-line*, resulting in a pre-verbal priming effect. Importantly, the results from an off-line relatedness test suggested that the differences in priming occurring between low- and high-value items cannot be explained by simple lexical priming (from object head noun to verb).

McRae et al. (submitted) already showed that verbs in isolation can be predicted on the basis of typical agents and patients. The current experiment shows that these kinds of predictions are indeed used during on-line sentence processing. Interestingly, in our sentences, the subject and object NPs were chosen to have as little associative relations to the main verb as possible without resulting in unnatural sounding or implausible sentences. Thus, the nouns that we used were not (proto-) typical agents or patients for the verbs.[7]

[4] In an earlier experiment we measured relatedness scores for a sentence part including the subject NP and the verb in relation to both the related and the control probe. In a CMLP experiment we measured the priming effects directly after these same sentence fragments. We then correlated the difference in rating scores with the priming effects (difference in RTs between related nad control probes), and indeed found a significant correlation of .31 (p = .025, 1-sided). So the greater the difference in off-line relatedness between the related and the control probe, the greater the on-line priming effect.

[5] The number of items adds up to 40 instead of 41, as we excluded one item because of scoring difficulties. This item is excluded in the relatedness test as well.

[6] Most of our probes matched our primes in transitivity. If the items where intransitive probes were used were excluded, the results did not change. Therefore, we assume that structural priming did not have a significant effect in our experiment.

[7] The fact that our agents and patients were not prototypical is based on intuitions. Currently we are running a paper-and-pencil generation task where we

The advantage for the parser in our study, as compared to the McRae et al. study, was that information on agents and patients could be combined. Probably one non-typical agent or patient in itself would not be enough to generate useful predictions about the verb, however, as our data show, the combination of two arguments can in some cases be enough to restrict the domain of upcoming verbs significantly.

Clearly, on the basis of our data we can only make tentative suggestions about how verbs can be predicted during on-line sentence processing. Our findings have to be replicated in a further experiment where all relevant factors are manipulated independently. Further studies should also focus on the partial effects of the different sentence parts, for example by testing for effects at different, strategically placed, points during the sentence.

6 Acknowledgements

The first author is funded by the Dutch Organization for Scientific Research (NWO) under grant #360-70-091 to Roelien Bastiaanse. We are grateful to our colleagues of the aphasia and experimental linguistics discussion groups and two anonymous reviewers for helpful comments on an earlier version of this paper.

References

Ahrens, K. (2003). Verbal integration: the interaction of participant roles and sentential argument structure. *Journal of Psycholinguistic Research, 32*, 497-516.

Altmann, G. T. M. & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition: International Journal of Cognitive Science, 73,* 247-264.

Berkum, J.J.M., van, Brown, C.M., Hagoort, P., Zwitserlood, P., & Kooijman, V. (2002). *Can people use discourse-level information to predict upcoming words in an unfolding sentence? Evidence from ERPs and self-paced reading.* Poster presented at the 8th International Conference of Cognitive Neuroscience (ICON-8), Porquerolles, France, September 9-15, 2002.

Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (in press). *Anticipating upcoming words in discourse: evidence from ERPs and reading times.*

Featherston, S. (2001). *Empty categories in sentence processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Friederici, A.D. & Frisch, S. (2000). Verb argument structure processing: the role of verb-specific and argument-specific information. *Journal of Memory and Language, 43*, 476-507.

Goede, D. de, Wester, F., Shapiro, L.P. , Swinney, D.A. ,& Bastiaanse, Y.R.M. (submitted). *The time course of verb processing in Dutch sentences.*

Hoeks, J.C.J., Stowe, L.A., & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19*, 59-73

Kamide, Y. (2004). *Filler-gap in the visual world.* Poster presented at the 10th Annual Conference on Architectures and Mechanisms of Language Processing (AMLaP-2004), Aix-en-Provence, France, September 16-18, 2004.

Kamide, Y., Altmann, G.T.M., & Haywood, S.L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133-156.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research, 29*, 627-645.

Kutas, M., & Hillyard, S.A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161-163.

McRae, K., Hare, M., Elman, J.L., & Ferretti, T.R. (submitted). *A basis for generating expectancies for verbs from nouns.*

Otten, M. & Berkum, .J.J.M. van. (2004). *Discourse-based lexical anticipation during language processing: Prediction or priming?* Poster presented at the annual meeting of the Cognitive Neuroscience Society (CNS-2004), San Francisco, April 18-20, 2004.

Salverda, A.P., Gennari, S., & Altmann, G. (2004). *Competing event-based interpretations mediate anticipatory eye movements in visual-world studies*. Poster presented at the 10th Annual Conference on Architectures and Mechanisms of Language Processing (AMLaP-2004), Aix-en-Provence, France, September 16-18, 2004.

Shapiro, L.P., Zurif, E., & Grimshaw, J. (1987). Sentence processing and the mental representation of verbs. *Cognition, 27*, 219-246.

present our main verbs and ask participants to generate prototypical agents and patients to confirm these intuitions.

The role of the verb during on-line spoken sentence comprehension

Dieuwke de Goede
Femke Wester
Roelien Bastiaanse
University of Groningen
Graduate School of Behavioral and Cognitive Neurosciences
Department of Linguistics, PO Box 716
9700 AS Groningen, The Netherlands

d.de.goede@let.rug.nl

David Swinney
University of California, San Diego
Department of Psychology
9500 Gilman Drive
La Jolla, CA 92093-0109

Lewis Shapiro
San Diego State University
Department of Communicative Disorders
5500 Campanile Drive
San Diego, CA□2182-1518

Abstract

The verb is the most central element in a sentence. Nevertheless, the exact role of the verb in the ongoing integration of various critical parts of the sentence as it unfolds in time is unknown. This paper reports the results of three Dutch Cross-Modal Lexical Priming experiments detailing the nature of activation of verbs during on-line sentence processing. The first two experiments show that in main clauses ending with a direct object, transitive matrix verbs remain active during the entire clause. Activation of the verb only dissipates upon encountering the conjunction, signalling a new clause. The third experiment shows that even when the final argument (direct object) is followed by an adjunct, continued activation of the verb is found until the end of the clause, so saturation of the argument structure does not result in verb de-activation.

1 Introduction

The verb is the core of a sentence: it expresses the event or activity that the sentence describes (by its proper meaning), it provides the number of possible persons or objects involved in the event (argument structure), and the links between these constituents and the verb itself (theta role assignment). Thus, verbs (and all the information we implicitly understand when we ’know’ a verb) provide a bridgework for nearly all aspects of sentential processing. While much research has demonstrated that verb information (like argument structure) plays a role in such ongoing processing, there remains a surprising paucity of evidence detailing the precise nature of how and when such information is employed.

The work presented in this paper is focused on filling part of this void, with the specific goal of determining the lifetime of a verb during the unfolding sentence, and trying to find how its linkage to important sentence constituents like arguments and adjuncts is reflected.

2 Verb meaning

There is an extensive literature on the semantics of verbs, and the effects of semantic complexity on on-line verb processing have been established frequently as well (e.g. Gennari & Poeppel, 2003). But the influence of verb meaning on processing at the sentence level, and the interaction between the meaning of the verb and the overall meaning of the sentence are less clear.

A couple of studies have investigated how during sentence processing the meaning of the verb in combination with all other sentential elements are integrated into more general representations of sentence meaning. With a sentence sorting task, Healy and Miller (1970) showed that verbs influence the judged meaning of a sentence to a greater extent than nouns (in this case: agents). However, when Gentner and France (1988) studied the differential effects of verbs and nouns on sentence interpretation, their conclusion was that if semantic strains force meaning adjustment (because the verb and the nouns do not converge into a natural interpretation), the verb is more often the locus of change than the noun is. According to these authors, verbs are more ‘mutable’ than nouns, probably because they are referential in nature (verbs have to convey relations or events that apply to the referents established by the nouns), and verbs are less semantically coherent, because they have more external links compared to nouns (external links allow context to influence interpretations, internal links cause stable interpretation).

The question that is raised here is whether the meaning of a verb is static or malleable when processed as part of a whole sentence. What has been found is that both additional semantic and structural information (argument structure) added to a verb can result in an importantly different interpretation of the message. Gentner (1981) combined simple verbs with additional semantic information (e.g. *give* + she *owed him money*), and found that this produced a structure identical to the meaning of another, more complex verb (e.g. *pay*).

Bencini and Goldberg (2000) used Miller's sorting task to disentangle the effects of the verb itself and the argument structure with which it is combined in a particular sentence. They found that argument structure patterns are directly associated with overall sentence meaning (e.g., *Anita threw the hammer* versus *Chris threw Linda the pencil*).

So, when an interpretation has to be established for a verb as part of a whole sentence, verbs seem to be very sensitive to merge with and adapt to other elements. It is not known how this affects the behaviour of the verb during moment-to-moment sentence processing.

3 Argument structure

It is well-established that verb-argument structure has direct consequences for the sentence processing system. Influences of argument structure have been found at very early processing stages, using a variety of experimental techniques (Mecklinger, Schriefers, Steinhauer, & Friederici, 1995; Pickering & Branigan, 1998; Shapiro, Zurif, & Grimshaw, 1987; Trueswell & Kim, 1998). These experiments have in common that they look for effects of argument structure directly after the verb itself. However, it is not known what happens afterwards, during the ongoing sentence, for example, when the direct object is encountered and a theta role has to be assigned. The verb is critical to the assignment of thematic roles to the arguments. But is it the case that the verb remains active until its arguments are encountered in the sentence, to fulfil the requirement that all arguments are assigned a thematic role?

There are a few clues from the literature. Examinations of verb processing in sentence context that focused on argument structure violations provide evidence that when an 'illegal' argument is encountered subsequent to the verb, the system immediately recognizes the error (Friederici & Frisch, 2000; Trueswell, Tanenhaus, & Kello, 1993). One interpretation of such effects is that the verb projects its possible arguments, setting up expectations or slots in which the arguments should appear. If those expectations are not met, the system is sent into error. Thus, there is no need to keep the verb active once it has projected its argument structure. However, another possibility is that the verb does remain active until its arguments are encountered and assigned an interpretation; as each argument is encountered and merged into the syntax, it is 'checked-off' from the verb's representation. Thus, if an incongruent argument is encountered, it is the verb that is the locus of the error response.

4 Experiments

We present three on-line Cross-Modal Lexical Priming experiments examining details of verb activation throughout the course of sentence comprehension in Dutch. In our materials, Dutch matrix clauses[1] are linked, via a conjunction, with embedded clauses which continue the sentence, to form a complex sentence. The first two experiments provide examination of verb activation at a number of different points up to and beyond the end of the matrix clause. The third experiment tries to assess whether the activation pattern of the verb is related to its argument structure.

4.1 Method

4.1.1 Participants

We tested 44 participants in Experiment 1, 60 in Experiment 2, and 48 in Experiment 3.

4.1.2 Paradigm

We used the Cross-Modal Lexical Priming task, a dual task in which participants listen to sentences and make a lexical decision to a visual probe presented at a particular point during each sentence. Facilitation in reaction times to a probe that is associatively related to a particular word in the sentence (prime) as compared to reaction times to a probe that is unrelated (but otherwise comparable to the related probe) is attributed to priming effects.

4.1.3 Materials

Sentences consisting of a matrix clause (SVO) followed by an embedded clause were aurally presented. In the first two experiments the matrix clause ended after the direct object. In Experiment 1 the direct object occurred directly after the verb (see 1), in Experiment 2 an adjunct preceded the direct object (see 2). The adjunct was inserted to extend the main clause, allowing the control probe point to be placed significantly later than in Experiment 1. In experiment 3, an adjunct was inserted after the Object Noun Phrase (see 3), to be

[1] In the current paper we ignore the fact that, according to some linguistic theories, the verb is not in its base position in Dutch matrix clauses (Koster, 1975). This issue is discussed in other work (e.g., De Goede, Wester, Shapiro, Swinney, & Bastiaanse, submitted).

able to test after the direct object. The adjuncts that were used were Adverbial Phrases of Time.

(1) De kleine jongens <u>imiteren</u> [1] hun fanatieke [2] rood-aangelopen voetbaltrainer, omdat [3] ze later allemaal profvoetballer willen worden.
The little boys <u>imitate</u> [1] their fanatical [2] red-faced soccer coach, because [3] they later all pro-soccer players want to become.

(2) De beschaafde mannen <u>imiteren</u> [1] regelmatig hun hysterisch [2] kijvende vrouwen [3], want [4] zo kunnen ze uiting geven aan hun frustratie zonder gewelddadig te worden.
The refined men i<u>mitate</u> [1] regularly their hysterical [2] cantankerous women [3], because [4] in-this-way can they express their frustration without violent to become.

(3) De kleine jongetjes <u>imiteren</u> [1] de fanatieke voetbaltrainer elke zaterda[2]gochtend weer [3], want ze willen later allemaal profvoetballer worden
The little boys <u>imitate</u> [1] the fanatical soccer-coach every Saturday[2]-morning again [3], because they want later all pro-soccer-player become.

The visual probes that were presented during the experimental sentences were verbs that were either associatively related to the finite verb (*nadoen = to copy*) or unrelated (*filmen = to film*) but matched to the related probe as good as possible on baseline lexical decision time, frequency, length and argument structure. Both probe types were pre-tested off-line for any possible inadvertent source of priming. The same prime - unrelated probe - related probe triads were used in all experiments.

Probes were presented at five different positions (see example sentences):

1. *verb probe point*: indicated as [1], placed directly after the verb, at the onset of the next word
2. *control probe point*: indicated as [2], presented at 700 ms after [1] in Experiment 1 and at 1500 ms after [1] in Experiment 2
3. *adjunct probe point:* measured in Experiment 3 only and indicated as [2], presented 700 ms after the onset of the first word of the adjunct
4. *end-of-clause probe point*: measured in Experiment 2 and 3 and indicated as [3], presented at the end of the clause
5. *conjunction probe point*: indicated as [3] in Experiment 1 and as [4] in Experiment 2, presented at the offset of the conjunction

In Experiment 1 and 3, 42 experimental sentences were used, in Experiment 2 there were 40 experimental sentences. In each experiment, an equal number of pseudo-experimental sentences (sentences with the same structure as the experimental sentences) were added and combined with non-words, to prevent correlation between sentence type and response type. In addition, 20 filler sentences of different structures (10 words, 10 non-words) and 15 yes/no comprehension questions were added (to encourage participants to pay attention to the spoken sentences).

A completely counterbalanced design was created to assure that all participants saw both related and control probes, and saw probes at all three probe points. Each participant was tested twice, on the same list, but with related and control probes shifted.

4.1.4 Procedures

The participants were tested individually in a sound-proof room with no visual distracters. The sentences were presented over headphones with an interval of 1500 ms. The probes were presented on a standard computer screen. The experimental software Tempo (developed at the University of California, San Diego, for running CMLP-studies), combined with a response box with two buttons, was used to present the items and register the accuracy and RTs of the responses. Each probe was presented for 300 ms and a response could be given within a 2000 ms interval from stimulus onset. Importantly, the sentences continued without interruption during visual presentation of the probe.

Participants were instructed to listen carefully to the sentences and to expect comprehension questions after some sentences. Questions were answered and lexical decisions were made by pressing the left button on the button box for no and non-word and the right button for yes and word. Participants were instructed to answer as quickly and accurately as possible.

4.2 Results Experiment 1 & 2

Participants were excluded from further analysis if their error score on the lexical decision task was greater than 10%, if the mean or SD of their reaction times (RTs) deviated from the overall mean or SD by more than 2.5 SD, or if less than 67% of the comprehension questions were answered correctly. Data from three participants were excluded in both experiments.

Error rates were low (1.4% and 1.8%, respectively) and equally distributed across related and control probes and across probe points. The exclusion of errors and outliers (all values deviating from the participants and item mean for the particular data point with more than 2.5 SD were excluded) resulted in 2.7 and 3.1 percent data loss, respectively.

The mean RTs for all probe points and probe types are presented in Table 1 (the values that are presented here and in the following tables are derived from the subject-analyses; the item-analyses revealed very similar data).

Probe Type	Probe Point			
	verb	*control*	*end-of-clause*	*conj.*
Experiment 1				
control	633	635	-	626
related	617	621	-	620
priming	16	14	-	6
Experiment 2				
control	663	671	668	666
related	662	657	654	672
priming	1	15	14	-6

Table 1: Mean reaction times to related and control probes for each probe point in Experiment 1 and 2.

The subject-based ANOVAs revealed a significant main effect of probe type in both experiments; overall, the related probes generated shorter RTs than the control probes (Exp 1: F1 (1,40) = 7.91, p = .008; Exp 2: F1 (1,56) = 4.61, p = .036). The item-based ANOVA was marginally significant in Experiment 1 (F2 (1,41) = 3.43, p = .07), but did not reach significance in Experiment 2 (F2 (1,39) = .98, p > .3).

Paired t-tests showed significant[2] faster responses to related than to control probes (priming) at the *verb* probe point in Experiment 1 (t1 (40) = 2.53, p = .008; t2 (41) = 1.75, p = .044), but not in Experiment 2 (t1(56) = .15, p > .4; t2 (39) = .29, p > .3). At the *control* probe point priming was found in both experiments (this effect was significant in the subject analysis (Exp 1: t1 (40) = 2.64, p = .006; Exp 2: t1 (56) = 2.49, p = .008), but in the item-analysis only a trend was found (Exp 1: t2 (41) = 1.40, p = .085; Exp 2: t2 (39) = 1.37, p = .090). Furthermore, priming was found at the *end-of-clause* probe point in Experiment 2 (t1 (56) = 2.08, p = .021; t2 (39) = 1.76, p = .043). Neither of the experiments, however, showed a priming effect at the *conjunction* probe point (Exp 1: t1 (40) = .81 p > .2; t2 (41) = .82, p > .2; Exp 2: t1 (56) = -.98, p > .15; t2 (39) = -.95, p > .15).

4.3 Interim conclusions

Experiment 1 and 2 converge on a pattern of activation of the verb at the *control probe points* (700 and 1500 ms after the actual occurrence of the verb in the sentence) and deactivation immediately following the *conjunction* linking the matrix to the second clause. The second experiment further shows that the verb is active at the *end of the clause*. The results for the *verb probe point*, where priming of the verb was expected directly after its occurrence, are less clear. Although significant facilitation of the related probe compared to the control probe was found in Experiment 1, Experiment 2 showed a null-effect at this probe position[3]. To explore this issue further, we tested at this position again in Experiment 3.

4.4 Results Experiment 3

The data were handled in the same way as in Experiment 1 and 2. Three participants were excluded from further analysis and exclusion of errors and outliers resulted in 3.0 percent data loss.

Probe Type	Probe Point		
	verb	*adjunct*	*end-of-clause*
control	721	723	712
related	697	706	693
priming	24	17	19

Table 2: Mean reaction times to related and control probes for each probe point in Experiment 3.

The RTs for this experiment are presented in table 2 and show faster responses to related probes than control probes at all probe points (F1 (1,44) = 24.35, p < .001; F2 (1,41) = 6.38, p = .016). First of all, directly after the *verb*, a significant priming effect is obtained, which indicates that a small adaptation in probe placement (the probe was now placed consistently at the onset of the first word

[2] As no inhibition effects were expected all t-tests are 1-tailed.

[3] Phonological assimilation made it impossible to place all probe points exactly at the onset of the next word. Post-hoc analyses showed that probes that were placed too early had a low probability to show faster RTs to related probes than to control probes, whereas the majority of probes that were presented exactly at the onset of the word following the verb showed facilitation for the related probe (□^{2} (1) = 17.4, p < .001).

following the verb or slightly later) resulted in stable priming effects at this probe point (t1 (44) = 3.08, p = .002; t2 (41) = 1.75, p = .044). Secondly, and more interestingly, activation of the verb was still evident after the final argument had been processed, 700 ms into the *adjunct* (t1 (44) = 2.35, p = .012; t2 (41) = 2.39, p = .011), as well as at the *end of the clause* (t1 (44) = 2.59, p = .007; t2 (41) = 1.77, p = .042).

5 General Discussion

In three experiments, Dutch complex sentences consisting of a matrix clause (SVO) followed by an embedded clause were aurally presented to participants. The Cross-Modal Lexical Priming (CMLP) paradigm was employed to examine the time course of activation of the verb during the unfolding sentence. The results of Experiment 1 and 2 revealed activation of the matrix verb once it was encountered, with activation continuing throughout the matrix clause. Activation appeared to dissipate when the conjunction, signaling a new clause, was encountered. Furthermore, Experiment 3 showed that the activation of the verb continued even after its final argument had been processed.

This continued activation of the verb contrasts sharply with the pattern that has been found for nouns: studies on anaphora and *wh*-movement have repeatedly shown that activation of anaphors and *wh*-traces rapidly degrades, according to Featherston (2001) the existing data converge on a figure of about 500 ms.

Why should verbs remain active across the unfolding of the sentence? One possibility centers on the critical role of the verb in providing the number and structural type of arguments, and assigning thematic roles to the arguments. Maybe the verb, after it has projected its argument structure, remains active until all arguments are encountered and interpretations have been assigned, which allows all arguments to be directly compared to the verb's representation. Our results are in line with this suggestion: we found activation of the verb during the occurrence of the direct object, the only argument that followed the verb in our sentences.

However, a strong reading of this account would predict activation of the verb only up to its final argument, and not after that. This is not what we found: in Experiment 3 we found activation of the verb during the adjunct, 700 ms after the direct object, although at this point in time it was unambiguously clear that the final argument had ended. Therefore, our results cannot be explained *solely* by a need of the verb to fill its argument structure. Probably other levels of conceptual structure play a role in keeping the verb active as well.

These studies suggest that verb interpretation is an ongoing process, which possibly only stops when a new clause and/or a new verb is encountered. During processing, each sentence constituent (the arguments, but possibly also other relevant information) is linked to the verb to zoom in to its specific interpretation. This interpretation is used to understand the event and to incrementally build up a proposition during sentence processing. The verb signals the main topic of the proposition, but the other sentence constituents are necessary to provide further details. At the end of a clause/sentence, when the verb has retrieved its meaning by linkage to the other sentential elements, a proposition is fully established.

Continued activation of a verb throughout a clause may be a critical aspect of sentence processing, in that material other than just verb arguments are critical to the interpretation of the conceptual information conveyed by the clause (and hence linked to the clausal verb in some fashion). It may turn out that adjunctive information may play a much more important role in sentence processing than has been heretofore considered. On intuition alone it is clear that sentences describe events with information about where, when, and why something happened, information often carried by adjuncts (as is formally captured by conceptual structure, e.g., Jackendoff, 2002, or mental models, e.g., Garnham, 2001; Johnson-Laird, 1983). Thus, it may not be surprising that verbs continue to be activated in their clauses beyond saturation of their arguments alone.

6 Suggestions for further research

On the basis of the previous discussion we can make a few predictions which might help to further disentangle the issues at stake. First of all, if argument structure is of importance, a different activation pattern would be predicted in the case of intransitive verbs. Using these types of verbs, extended activation of the verb would only be predicted in sentence constructions where the subject follows the verb, for example in a sentence like (4) and not in any other constructions like (5). The reason is that if the subject precedes the verb no arguments are to be expected after the verb.

(4) Elke zondagmorgen fietsen [1] de vrolijk giechelende [2] pubers in het wonderschone [3] groene park, omdat ....

Every Sunday-morning bike [1] the cheerfully giggling [2] adolescents in the wonderful [3] green park, because ....

(5) De vrolijk giechelende [1] pubers fietsen [2] in het wonderschone [3] groene park, omdat .....
The cheerfully giggling [1] adolescents bike [2] in the wonderful [3] green park, because....

So whereas in example (4) activation of the verb is predicted at probe point [2], for example 1000 ms after the verb, no activation is predicted at the same distance from the verb (probe point [3]) or at a position in the same NP (probe point [1]) in example (5).

Secondly, some predictions can be formulated on the basis of the assumption that the verb stays activated during a proposition: For example, a new verb clearly signals the start of a new proposition. Therefore verb activation should always dissipate when a new verb is introduced, even if this is not at the end of the clause, for example in a center-embedded construction, like (6). (This structure gives rise to another interesting questions: if, indeed the activation of the verb dissipates during the embedded clause, do we find re-activation when the direct object NP is encountered?)

(6) De kleine jongetjes imiteren [1], terwijl hun moeders het eten koken [2], hun fanatieke [3] voetbaltrainer, omdat ....
The little boys imitate [1], while their mothers the diner cook [2], their fanatical [3] soccer-coach, because ...

7 Acknowledgements

The first three authors are funded by the Dutch Organization for Scientific Research (NWO) under grant #360-70-090 to Roelien Bastiaanse. We are grateful to Edwin Maas for his invaluable help in setting up the first experiment and to Dirk-Bart den Ouden and Tracy Love for their many helpful comments and advice throughout various stages of the work. We also thank two anonymous reviewers for helpful comments on an earlier version of this paper.

References

Bencini, G.M.L., & Goldberg, A.E. (2000). The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language, 43*, 640-651.

Featherston, S. (2001). *Empty categories in sentence processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Friederici, A.D. & Frisch, S. (2000). Verb argument structure processing: the role of verb-specific and argument-specific information. *Journal of Memory and Language, 43*, 476-507.

Garnham, A. (2001). *Mental Models and the Interpretation of Anaphora*. Philadelphia, PA: Taylor and Francis.

Gennari, S. & Poeppel, D. (2003). Processing correlates of lexical semantic complexity. *Cognition, 89*, B27-B41.

Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory, 4,* 161-178.

Gentner, D. & France, I.M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.). *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence* (pp. 343-382). San Mateo, CA: Kaufmann.

Goede, D. de, Wester, F., Shapiro, L.P., & Swinney, D.A., & Bastiaanse, Y.R.M. (submitted). *The Time Course of Verb Processing in Dutch Sentences.*

Healy, A.F., & Miller, G.A. (1970). The verb as the main determinant of sentence meaning. *Psychonomic Science, 20*, 372.

Jackendoff, Ray S. *Foundations of Language. Brain, Meaning, Grammar, Evolution.* Oxford University Press 2002.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, UK: Cambridge University Press.

Shapiro, L.P., Zurif, E., & Grimshaw, J. (1987). Sentence processing and the mental representation of verbs. *Cognition, 27*, 219-246.

Trueswell, J.C. & Kim, A.E. (1998). How to prune a Garden Path by nipping it in the bud: fast priming of verb argument structure. *Journal of Memory and Language, 39*, 102-123.

Verbs, semantic classes and semantic roles in the ADESSE project

José M GARCÍA-MIGUEL **Francisco J ALBERTUZ**
ADESSE project
Facultade de Filoloxía e Tradución
University of Vigo
E-36200 Vigo, Spain
{gallego, albertuz}@uvigo.es

Abstract

This paper contains an overall description of ADESSE (http://webs.uvigo.es/adesse/), a project whose main goal is to manually provide definitions and information about semantic roles and semantic class membership for all the verbs in a syntactic database of nearly 160,000 clauses retrieved from a Spanish corpus of 1,5 million words.

1 Introduction

In this paper we outline the ADESSE (*Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*) project, developed at the University of Vigo. The goal of the project is to achieve a database with syntactic and semantic information about verbs and clauses from a corpus of Spanish. The main final outcome of ADESSE will be a corpus-based syntactic-semantic database including for each verb and each clausal construction in the corpus a pattern of arguments characterized in terms of syntactic function, phrase type, semantic features, and semantic role. This will be accompanied by absolute and relative frequencies for each constructional alternative.

The starting point is a syntactic database of contemporary Spanish (BDS)[1], containing the syntactic analysis of almost 160,000 clauses from a corpus of 1,5 million words. The main tables of the BDS contain a register for each clause, including general grammatical features of the clause (verb form, polarity, modality, voice, etc.) and related fields for any core syntactic argument. For each syntactic argument, the following features are offered:

- [SynFunc]Syntactic Function: Subject, Direct Object, Indirect Object, Oblique Object, Locative, Manner, Oblique Agent, Attribute
- [Agr/Clit] Verb agreement or object Clitic (if any)
- [SynCat] Syntactic Category, i.e. phrase type
- Preposition (if any)
- Animacy: Human, Concrete, Abstract, Propositional
- Definiteness
- Number

Table 1 shows an example from the BDS with some of the syntactic information that has been annotated, namely, the syntactic features that we consider more relevant for ADESSE.

<table>
<tr><td colspan="4"><i>Cuando estaba en la universidad me escribía canciones de amor</i> [TER:127]
‘When he was at the University, he used to write love songs for me’</td></tr>
<tr><td>SynFunc</td><td>Subj</td><td>DObj</td><td>IObj</td></tr>
<tr><td>Agr/Clit</td><td>3sg</td><td></td><td><i>me</i></td></tr>
<tr><td>SynCat</td><td></td><td>NP</td><td></td></tr>
<tr><td>Animacy</td><td>Human</td><td>Concrete</td><td>Human</td></tr>
</table>

Table 1. Basic syntactic information about a clause in the BDS

One of the most evident benefits of the BDS is that we can get detailed information about the syntactic constructions of the verbs registered in the corpus. However, the utility of the database would increase greatly if we could also add some semantic features, a task that is also being developed independently by other semantic annotation projects (Ellsworth et al 2004; Sgall et al 2004). So, the goal of ADESSE is to keep all the syntactic information from BDS, and to create new tables and fields for the introduction of relevant semantic information: semantic roles, verb senses, and verb classes.

Our theoretical background assumes the independence and semantic compatibility of verb meaning and construction meaning (García-Miguel 1995:24-25, Goldberg 1995). We think that the global meaning of a sentence combines the meaning of lexical items and the meaning of grammatical constructions in a non deterministic way, but in a process of partial compositionality (Langacker 2000:152). We also adhere to some tenets of frame semantics, and particularly to

[1] BDS is partly accessible at http://www.bds.usc.es/

some practices of the FrameNet project[2], although there are also some important differences that will be commented on below. Put simply, we think that the syntactic structure of the clause must be explained through semantics. The verb evokes a complex conceptual representation that includes some basic participants in a scene. The syntactic alternations with the same verb provide alternate construals of the scene focusing on different facets of the situations. With this problems in mind, ADESSE aims to become a data base for the empirical study of the interaction between verb meaning and construction meaning.

2 Verbs and Semantic Arguments

As it has been observed, each verb evokes a conceptual scenario which can be accounted for by describing the set of potential semantic arguments which that verb can be combined with. For example, the conceptual frame of *escribir* 'write' can be described by making use of four semantic roles: 0-Writer, 1-Text, 2-Recipient and 3-Topic. Though sometimes it is possible to express the whole set of semantic arguments, as in (a), syntactic constructions usually select a subset, profiling them in different ways and leaving the rest unexpressed, as in (b) or (c):

(a) *Juan* [0] *le escribió una carta* [1] *a su madre*
[2] *sobre sus recuerdos de infancia* [3]
‘John wrote a letter to his mother about his childhood remembrances’
(b) *Juan* [0] *escribió una carta* [1]
‘John wrote a letter’
(c) *Juan* [0] *le escribió a su madre*[2]
‘John wrote to his mother’

What definitively proves that syntax is not enough is that, sometimes, the same syntactic construction can be mapped with different configurations of semantic arguments. Compare examples (b) and (c) below, from the verb *sustituir* 'substitute, replace', [0-Agent / 1-Substituted (Old Entity) / 2-Substitute (New Entity)], where the syntactic pattern Subj DObj corresponds to two semantic schemas (0-1 and 2-1):

(a) *Rijkaard* [0] *sustituyó a Xavi*[1] *por Deco*[2]
‘Rijkaard replaced Xavi with Deco’
(b) *Rijkaard* [0] *sustituyó a Xavi*[1]
‘Rijkaard replaced Xavi’
(c) *Deco*[2] *sustituyó a Xavi* [1]
‘Deco replaced Xavi’

Finally, it is possible that (what is at first considered) one verb evokes, in different instances, frames corresponding to different semantic domains. For example, the verb *enseñar* admits uses as the following ones:

(a) *Ella* [0] *le* [2] *enseñaba su idioma* [1]
‘She taught him her language’
(b) *Ella* [0] *le* [2] *enseñaba las fotos* [1]
‘She was showing him the pictures’
(c) *Ella* [0] *enseñó al niño* [2] *a caminar* [1]
‘She taught the baby how to walk’

It seems clear that we must distinguish two frames, one corresponding to the domain of Cognition (examples a and c, roughly equivalent to English *teach*, despite the differences in syntactic construction) and the other to Perception (example b, English *show,* despite the fact that the constructions is similar to that in a). In cases such as this one, we need different sets of semantic roles for labelling verb arguments [0-Teacher, 1-Thing taught, 2-Learner vs. 0-Shower, 1-Thing shown, 2- Seer], so we postulate two different verb senses.

In order to account for these and other similar facts, the design of our database takes a structure, whose main tables and relations are depicted in Figure 1

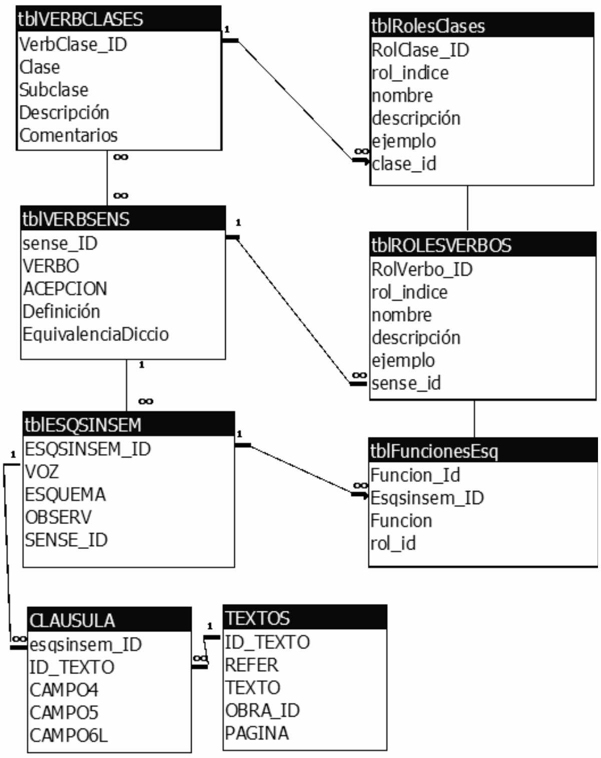


Figure 1: ADESSE database (partial) structure

Each record of the BDS (“Clausula” in fig. 1) is linked to a table of syntactic schemata (“tblEsq-

[2] See http://www.icsi.berkeley.edu/~framenet/ and, for an overview, Fillmore et al. (2003).

SinSem") where we map each syntactic function with a participant index (the equivalent of "0", "1", and "2" in the examples above). Each schema, in turn, is linked to a verb sense (tblVerbSens), associated with a set of participant roles, and ascribed to one or more semantic classes (tblVerbClases). The following sections explain the process in more detail.

3 Defining and unifying Verb Senses

Since our starting point is a database that contains very little semantic information, our first task has been to identify and define verb senses. This includes, among other things, a rough definition, a pointer to dictionary entries, and the splitting of a lemma into several verb senses when a unitary definition is not possible.

With respect to the distinction of verb senses, it must be remembered that our main interest is clause structure and not lexicology or lexicography, so we have not applied most of the criteria used in the lexicographical work. According to our theoretical background and our practical aims, we only distinguish verb senses when they are associated with different sets of semantic roles (see *enseñar* above). For example, the verb *escribir* has in ADESSE a single entry encompassing different subsets of a unique role set, despite the fact that some Spanish dictionaries distinguish up to three senses

Other lexical databases such as WordNet (Fellbaum 1998) follow a completely different way and admit a highly polysemic structure. That is, each possible group of synonyms ("synset") gives a new sense and then a new verb entry. So for example, WordNet 2.0 distinguishes 9 senses of *write*, 4 senses of *replace*, 13 senses of *show* and 2 senses of *teach*. ADESSE recognizes just one sense in each case for the Spanish equivalents of that English verbs.

Among the typical cases that do not imply different verb entries in ADESSE, one finds the following ones:

(a) *Constructional alternations,* whose meaning differences can be attributed rather to the constructional schema than to the verb. Under a single verb entry we can find voice alternations (active, middle, passive), causative/inchoative alternation, locative alternation, or some other rearrangement of arguments. In fact, the corpus recording of constructional alternations is the main goal of the ADESSE project.

(b) *Paradigmatic alternatives* inside an argument slot. Many verbs adjust their meaning depending on the nature of their more central arguments. For example, Spanish dictionaries distinguish about 15 senses of the verb *montar* 'mount', correlating with the nature of the thing mounted: a horse ('ride'), a concrete object ('assemble'), a business ('found', 'start'), an egg ('whip'), etc. Nevertheless, the schematic features of the argument structure do not vary very much and ADESSE contains just two senses of *montar*: 'ride' vs. 'assemble, set up'.

(c) *Metaphoric and metonymic uses* that can be extended or mapped from the basic sense of the verb. Nevertheless, although metaphoric uses do not suppose a new verb entry, they are identified and annotated in the corpus.

4 Verb Classes

In ADESSE each verb (in each sense) is given one (sometimes more) semantic class label(s). We use a hierarchical classification with two main levels: class and subclass. At the present we recognise 12 verb classes which reflect large semantic domains. Some classes can be grouped altogether into larger macroclasses, similar to some extent to Halliday's (2004) types of process:

MACROCLASS	CLASS	VERBS
1 Mental	11 Feeling	186
	12 Perception	72
	13 Cognition	122
2 Relation	21 Attribution	132
	22 Possession	117
3 Material Processes	31 Space	513
	32 Change	394
	33 Other facts	205
	35 Behavior	152
4 Communication		258
5 Existence		115
6 Causative and dispositive		57
TOTAL VERBS		

Table 2. *Top-level classes in ADESSE*

However, our basic and more useful category is subclass. Verb classes are therefore divided into 51 subclasses, associated with more concrete conceptual frames, each of which provides a (partially) specific set of semantic roles for labelling verb arguments (see below).

For example, the verb class Change splits into 5 subclasses as shown in Table 3:

SUBCLASS	VERBS
3200 General	14
3210 Creation	30
3220 Destruction-Consumption	35
3230 Modification	298
3231 Personal Care	17

Table 3. *Change subclasses in ADESSE*

In each verb class there is a General subclass including verbs with a more schematic content. For example, Change verbs such as *pintar* 'paint' or *cocinar* 'cook' are considered General Change verbs because they admit both Modification and Creation readings[3]. On the other hand, some subclasses are actually verbal groups inside a subclass, and identify more specific sets of verbs for further study. Thus Grooming or Personal Care verbs (3231), such as *lavar* 'wash' or *cepillar* 'brush', constitute a subtype of Modification verbs (3230), as reflected in the numerical index.

Unlike other verbal typologies, which use a fixed inventory of top-level categories, or which introduce the typology as the final outcome of a complete analysis of verbs, our classification is still provisional and its current structure represents working hypotheses about semantic organization that are always tested (and corrected, if necessary) for usefulness and empirical adequacy. As a point of departure, we have reviewed other semantic classifications, from the more lexically oriented (as WordNet) to the syntactical-semantic ones based on diathesis alternations (Levin 1993), though our premises fit better with proposals such as Dixon's (1991), Halliday's (2004) and FrameNet's (but see below). In fact, most of our clasess and subclasses are present in most classifications, but often with important differences in extension and hierarchical position. Some similarities between our system and WordNet high-level categories are evident.

Semantic verb classes in ADESSE are not empirically well-defined sets; rather, they represent generalizations over types of conceptual frames evoked by individual verbs in their specific instances, so problems of conceptual overlapping and fuzzy borders are expected, especially if, unlike WordNet, we are reluctant to divide verbal senses. Verbal meanings are multidimensional and highly flexible, and the classification of verbs is only possible by identifying the basic dimension(s) of meaning they profile and by keeping them apart from contextual influence. As an example, *frotar* 'rub' designates a Manner of Movement (without displacement, as *acunar* 'rock (to sleep)'), but it sems to profile a contact (as *tocar* 'touch') made by exerting force (as *presionar* or *pulsar* 'press') that can cause a modification/displacement of an entity (as *limpiar* 'clean'). Therefore, *frotar* has been classified as an Other Facts:Contact verb. Sometimes, however, verbs seem to equally profile more than one semantic dimension (and equally evoke more than one conceptual frame), so ADESSE allows multiple classification: *escribir* 'write' belongs to Change: Creation and Communication:General subclasses (as *crear* 'create' and *decir* 'say' respectively); *durar* 'last' is a verb of Existence:Time and also of Attribution:Value (as *tardar* 'delay' and *costar* 'cost' respectively), etc.

5 Semantic Roles: Between Verb Senses and Verb Classes

The identification and annotation of semantic roles is a fundamental task of the project, given that the basic goal is to document empirically the linking of syntactic functions and semantic roles. This goal should be achieved at any predefined level: semantic class, verb senses, syntactic schemata, and clauses of the corpus. In order to simplify a bit the manual process of annotation and to achieve a greater coherence within the database, we assume that each level inherits by default the semantic information established in the higher levels; that is, in principle, we do not annotate each clause in the corpus, but the syntactic schemas that they instantiate. Syntactic schemas, in turn, point to roles that are defined for each verb sense. And verb participant roles can inherit features and labels from class-defined participant roles. In any case, we account for the possibility that each lower level contradicts or increments the information inherited from the higher levels.

First, each conceptual (sub)class is associated with a set of semantic roles prototypical for the cognitive domain denoted by the verbs belonging to it. Role labels are created by aiming at specificity (with class-specific labels) and transparency (descriptive adequation), trying to use, as far as possible, widely used traditional labels. Here are the role labels associated with some classes:

Change:Modification:
A0:Agent; A1:Affected

Communication:
A1:Sayer; A2:Message; A3:Addressee; A4:Topic

Feeling:
A1:Experiencer; A2:Stimulus

Possession:Belonging:
A1:Possessor; A2:Possessed

Space:Displacement
A0:Causer; A1:Theme; A2:Source; A3:Goal

[3] Compare *No había cocinado espárragos desde que ella llegó a casa* 'She had not cooked asparagus since she had arrived home' [BAIRES:493, 21] with *Podríamos pasar las veladas* […] *cocinando "escudellas del Ampurdán"* 'We could spend the evenings […] cooking *escudellas del Ampurdán*' [a typical Catalonian dish] [AYER:24, 5].

Secondly, each verb entry is associated with a set of semantic roles embracing any possible core participant in the scenes designated by the verb in any syntactic schema (see above examples with *escribir, sustituir,* and *enseñar*). In general, a set of explicit inheritance relations makes a verb inherit by default the roles considered basic for the class to which it belongs, although some verbs need some additional arguments in order to account for any syntactic construction with such verbs. For example, the verb *sustituir*, a member of the class Other facts:Substitution, inherits a set of roles that is common to other verbs of the same class (*reemplazar, cambiar2, suplir*, etc):

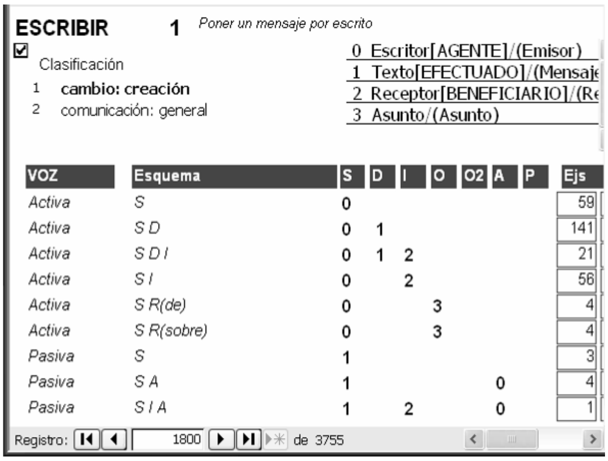
	A0	A1	A2
SUBSTITUTION	Agent	Substituted	Substitute
Sustituir	Agent	Substituted	Substitute

However, verb-specific role labels are used whenever there is a total or partial mismatch between a verb argument and class-specific role labels. For example, the verb *escribir* ‘write’ is both a Creation verb and a Communication verb. Its argument roles are inherited from Creation (Agent – Effected – Beneficiary) and from Communication (Sayer – Message – Addressee – Topic); but for the sake of clarity, the first two participants are labelled as Writer and Text.

	A1	A2	A3
COMMUNICATION	Sayer	Message	Addressee
Escribir	Writer	Text	Recipient

Third, the syntactic constructions of each verb are annotated simply with a pointer from each syntactic argument to one of the roles defined for the verb entry. This pointer allows us to trace the correspondences between arguments of different syntactic schemas (the pointer being identical for the equivalent arguments of diathesis alternations such as active / passive, causative / inchoative and so on). For example, in Figure 2, both the active voice object [D] and the passive voice subject [S] get the pointer “1”, indicating the Text written[4]. Given that syntactic functions are linked to a pointer, we could change the labels or the details of the classification without touching the essential aspects of the diathesis alternations.

Multiplying syntactic schemas by verb senses, we get about 12500 syntactic-semantic schemas that constitute the main target of our annotation. Given that each clause of the corpus is being linked to a syntactic-semantic pattern, we think that this strategy will allow us to characterize semantically the 159,000 clauses of the corpus in a relatively short time. This way, each clause is receiving an annotation similar to Table 4, which expands the example in Table 1.



ESCRIBIR 1 *Poner un mensaje por escrito*

☑ Clasificación

1 **cambio: creación**
2 comunicación: general

0 Escritor[AGENTE]/(Emisor)
1 Texto[EFECTUADO]/(Mensaje
2 Receptor[BENEFICIARIO]/(Re
3 Asunto/(Asunto)

VOZ	Esquema	S	D	I	O	O2	A	P	Ejs
Activa	*S*	0							59
Activa	*S D*	0	1						141
Activa	*S D I*	0	1	2					21
Activa	*S I*	0		2					56
Activa	*S R(de)*	0			3				4
Activa	*S R(sobre)*	0			3				4
Pasiva	*S*	1							3
Pasiva	*S A*	1					0		4
Pasiva	*S I A*	1		2			0		1

Registro: 1800 de 3755

Figure 2. Patterns of *escribir* in ADESSE

Me escribía canciones de amor [TER:127] ‘He used to write love songs for me’			
Escribir	Writer	Text	Recipient
CREATION	Agent	Effected	Benefactive
COMMUNIC.	Sayer	Message	Addressee
SynFunct	Subj	DObj	IObj
Agr/Clit	3sg		*me*
SynCat		NP	
Animacy	Human	Concrete	Human

Table 4. Syntactic and semantic annotation of arguments in a clause of BDS+ADESSE

6 Comparing with FrameNet

Our classification has a clear conceptual basis, which makes it very similar in some respects to FrameNet. Nevertheless, there are some important differences, beginning with the fact that we use a syntactically analyzed corpus to semantically annotate all and only the clauses in the corpus, not a set of selected sentences that illustrates frames and lexical units.[5]

Moreover, in FrameNet, the basic unit is obviously the *Frame*, so that Frame Elements and Lexical Units are defined in relation to the frame they belong to. In ADESSE, by contrast, the basic unit is the verb. Classes and subclasses represent generalizations over argument configurations in an attempt to get a set of role labels applicable by default to the verbs of the same class.

On the other hand, and more relevant in practice, ADESSE classes and subclasses are much

[4] This strategy has many similarities with PropBank annotation procedure (Kingsbury-Palmer 2002).

[5] In this respect, our goal is similar to that of PropBank and SALSA (Ellsworth et al 2004).

more schematic than frames in FrameNet[6]. This appears to be self-evident if we compare our 52 classes with the more than 300 frames containing verbs. Therefore, in ADESSE verbs such as *ver* 'see' and *mirar* 'look at' or *oír* 'hear' and *escuchar* 'listen' are included in the Perception class, disregarding semantic features as intentionality or attention which justify the FrameNet distinction between Perception_Experience and Perception_Active frames.

In line with our theoretical background, in ADESSE we try to keep apart verb meaning and construction meaning, and consequently we do not delimit verb senses simply on the basis of constructional alternations. FrameNet dissociates in different frames, for example, any verb participating in the locative alternation. Therefore, *load* in *John loaded the wagon with hay* is assigned to the frame Filling, whereas *load* in *Betty load the stuff in the car* is included in the frame Placing. By contrast, ADESSE unifies the spatial senses of *cargar* 'load' under just one verb sense under the class Localization. The meaning differences observed as a consequence of the 'locative alternation' are attributed to the meaning of the respective argument-structure constructions (in line with Goldberg 1995).

Moreover, ADESSE classes allow a variable degree of correspondence between a verb's argument structure and the pattern of participant roles prototypical for the class it belongs. For example, *mentir* 'lie' and *callar* 'be silent' are Communication verbs although *mentir* does not combine with a Message nor *callar* with a Recipient.

Last, apart from class-specific role labels, ADESSE can use verb-specific role labels. By default, verb-specific role-labels are inherited from class-specific role-labels, even though a verb can have a set of roles partly different from the class to which it is ascribed. This is the case of the verb *escribir* 'write' commented above. The use of verb-specific role-labels does away with the need to create new frames whenever the class or subclass is too wide.

7 Conclusion

At the time of writing this paper, the ADESSE project contains a provisional semantic classification of about 1700 verb senses, and an index of semantic role for each argument of about 4000 syntactic-semantic schemas, which correspond to more than 50000 clauses of the corpus. There is a lot of work to be done, but we aim to achieve a useful database for descriptive studies of the interaction between verbs and constructions in Spanish. So that we can obtain, for example, the diathesis alternations for any verb, the syntactic realizations of a participant role, or the syntactic constructions for a semantic domain (and vice versa).

8 Acknowledgements

ADESSE is being supported by the University of Vigo, the Spanish Ministry of Science and Technology (BFF2002-01197) and the Galician Autonomous Government (PGIDIT03PXIC30201PN). We also acknowledge a lot of people in Vigo and Santiago that have contributed to the building of the BDS for many years.

References

Dixon, Robert M. W. 1991. *A New Approach to English Grammar, on Semantic Principles*, Oxford University Press, Oxford.

Ellsworth, M. / K. Erk / P. Kingsbury / S. Padó. 2004. PropBank, SALSA, and FrameNet: How Design Determines Product. In *Proceedings of LREC-2004*, Lisbon.

Fellbaum, Christiane. 1998. "A Semantic Network of English Verbs". In *WordNet: An Electronic Lexical Database*, Fellbaum, Christiane, ed., pages 69-104, MIT Press, Cambridge (MA).

Fillmore, C.J. / C. Johnson / M. Petruck. 2003. Background to FrameNet. In *International Journal of Lexicography,* 16/3: 235-250.

García-Miguel, José M. 1995. *Transitividad y complementación preposicional en español.* Universidade de Santiago de Compostela.

Goldberg, Adele. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago

Halliday, M.A.K. 2004. *An Introduction to functional grammar*. E. Arnold, London (3rd edition)

Kingsbury, P. and M. Palmer. 2002. From TreeBank to PropBank. In *Proceedings of LREC-2002*. Las Palmas.

Langacker, Ronald. 2000. *Language and Conceptualization.* Mouton de Gruyter, Berlin.

Levin, Beth. 1993. *English Verb Classes and Alternations: a Preliminary Investigation.* University of Chicago Press, Chicago.

Sgall, P. / J. Panevová / E. Hajičová. 2004. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In *Proceedings of HLT-NAACL-2004*. Boston.

[6] Nevertheless, FrameNet has frames at different levels of schematicity. More schematic frames, inherited or used by more specific ones, are most similar to ADESSE classes and subclasses. In fact, FrameNet I grouped specific frames into semantic 'domains'.

Token-level Disambiguation of VerbNet classes

Roxana Girju, Dan Roth, and Mark Sammons
Computer Science Department
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL 61801
USA,
{girju, danr, mssammon}@cs.uiuc.edu

Abstract

The automatic disambiguation of verbs in domain independent text becomes more and more important for applications such as Machine Translation, Text Summarization, and Question Answering, mainly because verbs play a key factor in the syntactic and semantic interpretation of sentences. In this paper we present a system for the automatic classification of token verbs in context based on VerbNet classes. A supervised machine learning classifier is trained and tested on a portion of PropBank using a set of lexical and syntactic features.

1 Introduction

The automatic disambiguation of verbs in domain independent text becomes more and more important for applications such as Machine Translation, Text Summarization, and Question Answering (QA). Although a lot of work has been done on verb classification ((Palmer, 2000), (im Walde, 2000), (Merlo and Stevenson, 2001), (Lapata and Brew, 2004)), the focus was more on verb *types* than *tokens* [1]. While verb types are of great linguistic interest, from a natural language processing (NLP) perspective, token-level disambiguation is the challenge. The semantic classification of a given verb in context is a key factor in the performance of *semantic parsers* due to the verb's relevance to argument structure. Such a semantic parser enriched with verb semantic information can be then employed, for example, in QA proving.

One of the difficulties of the token-level verb disambiguation is that it requires massive text collections annotated with verb semantic information. However, the development of large semantically annotated corpora, such as Penn Treebank2 ((Marcus, 1994)) and, more recently PropBank ((Kingsbury et al., 2002)) and FrameNet ((Baker et al., 1998)), as well as semantic lexicons, such as VerbNet ((Kipper et al., 2000)) make the task possible.

In this paper we focus on *token-level verb classification*, i.e. for each occurrence of a particular verb in a corpus, we label it with the corresponding VerbNet semantic class. Then, a classifier is trained based on the extracted contextual information and tested on a set of unseen verb instances. This supervised model takes as input a portion of the PropBank corpus and a set of lexical and syntactic features that successfully contribute to identifying the corresponding VerbNet semantic verb classes in context.

2 The Data

We rely on the semantic information provided by VerbNet and PropBank. VerbNet is a hierarchical lexicon of over 4,100 verbs organized into classes according to Levin's classification ((Levin, 1993)). In order to maintain the semantic and syntactic coherence of the lexicon's members for each class, VerbNet extends and refines the original Levin's classes with 74 new subclasses.

PropBank is a semantically annotated version of the Wall Street Journal portion of the Penn Treebank. The main goal of PropBank is to provide consistent semantic role labels across different syntactic realizations of the same verb. The annotation captures predicate-argument structures based on sense tags for polysemous verbs (called *rolesets*) and semantic role labels for each argument of the verb, as shown in the following example:

(1) "[Mary]/ARG0 left/*leave.01* [the room]/ARG1."

Here ARG0 represents the *leaver*, ARG1 the *thing left*, and *leave.01* the roleset.

[1] A token is an individual occurrence of a word. A type is a category, for which a token is an instance. Thus, verb type refers to the verb in general, whereas verb token is the usage of the verb in a particular sentence.

For each verb, the PropBank descriptions provide several *rolesets* representing coarse grained verb senses. This sense inventory is based primarily on *usages* of a verb and might have different argument structures or different syntactic alternations for each usage. The set of all possible rolesets of a verb is captured in the verb's *frame*. Besides listing the rolesets, the PropBank frames provide a mapping between them and the possible VerbNet classes a roleset can have. However, not all PropBank verbs have been disambiguated (e.g., assigned rolesets). Moreover, when disambiguated, not all rolesets are associated with a single VerbNet class. Some verb rolesets have a one-to-one mapping to VerbNet classes, while others can map to more than one class, the appropriate selection being provided by the context. Furthermore, some verb rolesets have no VerbNet class associated with them at all. Table 1 shows examples of possible <roleset - VerbNet class> mappings in PropBank.

Verb	**Rolesets**	**VerbNet classes**
say	say.01	37.7
assert	assert.01	48.1.2
	assert.02	29.5
	assert.03	29.5
bow	bow.01	40.3.3
	bow.02	47.3, 47.6, 50
	bow.03	-
accomplish	accomplish.01	no class provided

Table 1: Examples of mappings between rolesets and VerbNet classes in the PropBank frames.

As our approach is supervised, we needed a corpus in which each verb instance is already disambiguated and annotated with the corresponding VerbNet class in context. Thus, by using the mapping <rolesets - VerbNet classes> provided by the PropBank frames, we replaced each instance of a verb's roleset in the corpus with its corresponding VerbNet class. However, we selected only those instances for which the frames provided a one to one mapping. From the 4,653 rolesets in the PropBank frames, 339 mapped to multiple classes, 1,592 mapped to only one class, 2,180 didn't map to any of the VerbNet classes ("-"), and 542 didn't list any mapping. The total number of unique VerbNet classes provided in the frames is 221, from which only 206 were found in the one-to-one mapping to rolesets for a total set of 2,756 unique verbs[2]. It is crucial to note that the selection of only those verb rolesets that mapped to only one class does not make the task trivial, as each verb may still have several distinct rolesets and can, therefore map to more than one verb class. The only disadvantage is that it reduces the corpus coverage to only some of the verb instances in the text collection, those which have a one-to-one mapping.

For this evaluation, we used the data provided by the CoNLL-2004 *semantic role labeling* shared task ((Carreras and Màrquez, 2004)) which consists of the sections 15-18, 20, and 21 of the February 2004 release of the PropBank corpus.

Based on the one-to-one mapping generated from the PropBank frames, we built two corpora. They contain instances that could be found at least once (corpus A), and at least 10 times (corpus B) respectively in the CoNLL collection.

Table 2 shows the total number of unique verbs, unique rolesets, unique VerbNet classes, and the total number of instances of one-to-one mappings in each corpus. As shown in the table, the number of verbs with a frequency of occurrence less than 10 is small (row "Corpus A - Corpus B").

3 The Model

Given a verb in its sentential context <verb, sentence>, the goal is to develop procedures for the automatic labeling of the VerbNet semantic class it encodes. The semantic class derives from the lexical, syntactic, and semantic features of each verb token.

The semantic classification of verbs can be formulated as a learning problem, and thus benefit from the theoretical foundation and experience gained with various learning paradigms. This is a multi-class classification problem since the output can be one of a given set of semantic verb classes. We cast this as a supervised learning problem where input/output pairs are available as training data.

An important first step is to map the context information of each verb into feature vectors. We define with $\mathbf{x}_i$ the feature vector of an instance i and let X be the space of all instances;

[2] These statistics were computed on the February 2004 release of PropBank.

	No. unique verbs	No. unique rolesets	No. unique classes	No. instances
Corpus A	870	944	177	12,431
Corpus B	748	808	106	12,158
Corpus A - Corpus B	138	136	71	273

Table 2: List of unique verbs, rolesets, classes, and total number of instances in each corpus. The last row shows the number of verbs with a frequency of occurrence less than 10 in these corpora.

i.e. $\mathbf{x}_i \in X$.

Let T be the training set of examples or instances $T = (< \mathbf{x}_1 c_1 >, < \mathbf{x}_2 c_2 >, ..., < \mathbf{x}_l c_l >) \subseteq (X \times S)^l$ where l is the number of examples $\mathbf{x}$, each of which is accompanied by its semantic class label c. The problem is to decide which class c to assign to a new, unseen example $\mathbf{x}_{l+1}$. In order to classify a given set of examples (members of X), one needs to map the observed instance into a feature-based representation that encodes information that supports generalization.

3.1 SNoW Learning Architecture

We use the SNoW learning architecture ((Roth, 1998; Carlson et al., 1999)) as our classifier. SNoW is a very efficient multi-class classifier that is specifically tailored for large scale learning tasks in terms of both number of examples and number of features, and has been used successfully in a range of classification problems. SNoW is a linear classifier that allows several update rules to be used, including variations of Perceptron, Naive Bayes, and Winnow. We use a variation of the Winnow multiplicative update rule (Littlestone, 1988), which best addresses the high dimensionality and sparsity issues in NLP data. One of the important improvements SNoW incorporates over the basic Winnow update rule is a regularization term, which has the effect of trying to separate the data with a thick separator (large margin) ((Grove and Roth, 2001)). For the experiments described here we use this regularization with a fixed parameter.

In the current classification task, each verb can potentially be mapped to a large number of verb classes, making this a very hard multiclass classification problem. In practice, though, the number of effective candidates is much smaller. As mentioned in section 2, a verb can map to several verb classes, but not to all of them, depending on its roleset. For example, the verb "*assert*" has associated as potential verb classes the following: "*{48.1.2, 29.5}*". Analyzing the corpus, we can determine the effective candidate set for each verb. We make use of the *sequential model* incorporated within SNoW ((Even-Zohar and Roth, 2001)) to restrict the set of candidates to this effective set, both in training and test. This makes the multiclass classification more tractable. We compare the sequential model with the *flat model*, in which we consider as potential classes of a verb all those that occur in the corpus.

4 Feature Space

So far, we have identified and experimented with the following **features**:

a). Word feature is the lexical form of each word in a window of size three surrounding the target verb.

b). Part-of-Speech tag (POS) feature is the POS tag of each word in a window of size three surrounding the target verb.

c). Chunk feature identifies the shallow-parse phrase type of each word in a window of size three.

d). Word & POS tag feature is the conjunction of each word and its POS tag for each word within a window of size three of the target verb.

e). Named entity feature shows the named entity associated with a particular word, for each word within a window of size three of the target verb.

f). Bigram feature. We generate bigrams of POS-tags, words, and POS-word combinations within a window of size three of the target verb. We also generate bigrams of the conjunctions described in feature d) above, also within a window of size three.

5 Experimental Results

Each classifier was trained and tested using a 10-fold cross validation on each corpus described in section 2. As a baseline we simply assign the most common class in each corpus to every instance in the test data, ignoring context and any form of prior information. We do this per verb and then take the average over all verbs in each corpus. The system's performance for each corpus considered is presented in Table 3.

The high accuracy results obtained by the sequential model and especially by the baseline on each corpus are explained by the fact that many of the verbs in these corpora are monosemous, eg. they were labeled with only one verb class. Specifically, 95.5% (corpus A), and respectively 96% (corpus B) of the verbs mapped to only one class. For example, the verb "*assert*" presented in Table 1 occurred only with the senses of "*assert.02*" and "*assert.03*" which generated a mapping to only one verb class: "*29.5*".

In order to see how well does the model work for ambiguous cases, we performed the same experiments only on those subsets of the corpora that contained ambiguous verbs. These verbs have an ambiguity degree of only 2. There were 39 unique verbs, 59 unique verb classes, and 972 instances in the ambiguous subcorpus A (28, 39, and 822 respectively in the ambiguous subcorpus B). The system's performance and the baseline values are shown in Table 4.

As expected, the baseline dropped considerably. The sequential model had an improvement of 6.71% (ambiguous subcorpus A) and 6.36% (ambiguous subcorpus B) over the baseline, compared with 2.05% (corpus A) and 2.1% (corpus B) in the previous set of experiments. This shows that the classifier also works well on ambiguous cases.

6 Related Work

Although most approaches in automatic verb labeling define the problem as a learning task, they differ in various respects. These include the inventory of classes used, learning models employed in generating semantic classes, and specific information about the verb (eg, type vs. token-level, monosemous vs. polysemous, etc.).

(Merlo and Stevenson, 2001) use a decision tree learner based on a set of grammatical features to classify verbs into three semantic classes. These sets are abstractions of Levin's classes, thus providing a more coarse grained classification. The main assumption is that differences in thematic roles uniquely identify semantic classes even though they share the same syntactic frame. The approach was tested only on a set of 59 manually selected Levin verbs. The learner achieves an accuracy of about 69%.

Schulte im Walde (2000) focuses on subcategorization frames and selectional restrictions employed in an unsupervised learning model. The approach is evaluated on 153 Levin verbs, out of which only 53 could map to more than one class. She obtained a precision of 61% and a recall of 36%.

(Lapata and Brew, 2004) provide a statistical model of verb class ambiguity by generating preferences for ambiguous verbs without the use of a disambiguated corpus. Additionally, they show that these preferences together with shallow contextual information can help a verb sense disambiguator. They restrict their model by focusing only on ambiguous verbs that are encoded by a set of five syntactic frames. Our model is more general in the sense that it takes general contextual features into account without limiting the coverage to a number of verbs encoded by a predefined set of syntactic frames. Moreover, Lapata & Brew's model is not context-sensitive. Thus, it cannot derive class rankings tailored to specific verbs as our model does. The results they report are generated per frame.

In Table 5 we summarize these approaches and compare them with our model based on various classification criteria.

7 Conclusion

In this paper we presented a supervised, token-level approach to the automatic classification of verbs into VerbNet semantic classes. We relied on a subset of a state-of-the-art semantically annotated corpus, PropBank, on which we trained a classifier based on lexical and syntactic features. The main contributions of the paper are:

- We treat VerbNet mappings as a sense tagging task (as (Lapata and Brew, 2004), but on a different data set),
- We get good results on a task based essentially on syntactic alternations without reference to full syntactic parses.

The results obtained are very promising. Moreover, the experiments performed suggested more future work:

- We used only shallow lexical and syntactic features to capture contextual information. We intend to extend the feature vector to capture semantic information. We will also investigate the use of full parses for such task.

Corpus	Model	System's accuracy [%]	Baseline [%]
Corpus A	**Sequential model**	98.47	96.42
	Flat model	38.24	
Corpus B	**Sequential model**	98.66	96.56
	Flat model	40.04	

Table 3: System's performance obtained for each experiment on each corpus. 95.5% (corpus A), and respectively 96% (corpus B) of the verbs didn't have ambiguity at all. This is what explains the high values for the Baseline.

Corpus	Model	System's accuracy [%]	Baseline [%]
Ambiguous subcorpus A	**Sequential model**	80.66	73.95
	Flat model	45.37	
Ambiguous subcorpus B	**Sequential model**	79.50	73.14
	Flat model	48.72	

Table 4: System's performance obtained for each experiment on the ambiguous corpora subsets.

- Many of the verb instances in corpora considered were monosemous (eg, as rolesets, and thus as verb classes), thus generating high accuracy values for the sequential model and baseline. We will also test the model on text collections with more polysemous instances.
- Currently, our approach is not capable of labeling unseen verbs in context. In this respect, we intend to extend the feature vector with rich local and global contextual information that would help guessing a verb class in context.
- We intend to compare our approach to similar ones used for the sense disambiguation of polysemous verbs as defined by the Senseval2 (Cotton et al., 2001) and Senseval3 (Mihalcea and Edmonds, 2004) tasks.

8 Acknowledgements

We would like to thank the anonymous reviewers for their useful comments. This research was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) grant, and in part by the Office of Naval Research (ONR)'s TRECC grant.

References

C. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLLING-ACL, Canada.*

A. Carlson, C. Cumby, J. Rosen, and D. Roth. 1999. The SNoW learning architecture. Technical Report UIUCDCS-R-99-2101, UIUC Computer Science Department, May.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In Hwee Tou Ng and Ellen Riloff, editors, *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning CoNLL-2004*, pages 89–97, Boston, Massachusetts, USA, May 6 - May 7. Association for Computational Linguistics.

Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer, editors. 2001. *SENSEVAL2: Second International Workshop on Evaluating Word Sense Disambiguation Systems.* Association for Computational Linguistics, Toulouse, France.

Y. Even-Zohar and D. Roth. 2001. A sequential model for multi class classification. In *EMNLP-2001, the SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 10–19.

Adam J. Grove and Dan Roth. 2001. Linear concepts and hidden variables. *Machine Learning*, 421/2:123–141.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behavior. *Proceedings of COLING-2000.*

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the penn treebank. *Proceedings of the Human Language Technology Conference HLT 2002, CA.*

Related work	**Verb Class inventory**	**Features**	**Models**	**Type/ Token-level**	**Monosemous/ /Polysemous**	**Performance**	**Focus**
Merlo& Stevenson	3 (Levin)	5 features with freq. counts	C5.0	type	monosemous	69.8%	classify verbs in 3 semantic classes
Schulte im Walde	30 (Levin)	verb freq. with subcat. frames	iterative clustering	type	polysemous	61%	discover Levin classes from corpora
Lapata & Brew	∗ (Levin)	informative priors; contextual information syntactic frame information;	Naive Bayes	type (token only for class disambig. task)	polysemous	87.8% (transitive frame only)	compute preferences for ambiguous verbs and test to see if they can help disambiguating verbs in context
Our model	(VerbNet) 177 (corpus A) 106 (corpus B)	contextual information;	SNoW	token	(some) polysemous	98.47% (corpus A) 98.66% (corpus B)	disambiguate verbs in context

Table 5: Comparison with previous work. "∗" means "not provided".

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. *Proceedings of AAAI-2000 Seventeenth National Conference on Artificial Intelligence, Austin, TX.*

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using information priors. *Computational Linguistics*, 301:45–73.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago.

N. Littlestone. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318.

Mitchell Marcus. 1994. The penn treebank: A revised corpus design for extracting predicate-argument structure. *Proceedings of the ARPA Human Language Technology Workshop, Princeton, NJ.*

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb-classification based on statistical distribution of argument structure. *Computational Linguistics*, 273:373–408.

Rada Mihalcea and Phil Edmonds, editors. 2004. *SENSEVAL3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Association for Computational Linguistics, Barcelona, Spain.

Martha Palmer. 2000. Consistent criteria for sense distinctions. *Computers and the Humanities*, 341-2:217–222.

Dan Roth. 1998. Learning to resolve natural language ambiguities: A unified approach. In *Proceedings of the American Association of Artificial Intelligence.*

The Role of Inflectional Morphology in Syntactic Parsing

Patrycja Jabłońska
Faculty of Humanities,
University of Tromsø/CASTL,
9037 Tromsø,
Norway,
patrjabl@yahoo.com

Svetoslav Marinov
IKI, Skövde University &
Faculty of Arts, Göteborg University,
405 30 Göteborg,
Sweden,
svetoslav.marinov@his.se

Abstract

Inflectional morphology is often considered a burden to the language learner and it has never found a place in syntactic parsing. In this paper we show how it constrains verbs' argument structure and present an approach to employing these restrictions in syntactic parsing of Polish.

1 Introduction

The present work assumes a neo-constructionist approach to argument structure (cf. *inter alia* (Arad, 1998), (Marantz, 1997),(Marantz, 2003), (Borer, 2003)) whereby the number and type of arguments associated with a particular root is not a lexical property of this root, but rather results from embedding the root in a given syntactic configuration. In view of a considerable flexibility w.r.t. argument structure configurations displayed by most of the roots in languages like English, this stand seems to be a conceptual advantage over 'rich' lexical entries for verbs supplemented by so-called 'linking rules' (cf. *inter alia* (Levin and Rappaport, 1995)).

On the other hand, there is a limit to the flexibility that must be incorporated into any account. The purpose of this paper is twofold. Firstly, we will argue here that some languages, which have morphological means at their disposal, exploit the system of *prima facie* redundant inflectional morphology to restrict argument structure flexibility. We take Polish as an example. Secondly, we show how such an approach to argument structure can be incorporated into NLP tools (e.g. a parser) and what are its advantages. For this purpose we use the multiparadigm programming language Oz (van Roy and Haridi, 2004) which offers concurrency and constraint handling. We illustrate this with a simple parsing system for Polish.

2 Conjugation class suffixes as verbalizers

Polish verbs are characterized by the presence of a suffix (usually vocalic) intervening between the root and a Tense/Agreement morpheme. This suffix defines a conjugation class of a given root. In the present work we concentrate on four classes only (presented in (1)), refraining from making any claims about other classes.

(1) a. **-aj- class**: *czyt-a-ć* ('read'), *mrug-a-ć* ('wink'), *śpiew-a-ć* ('sing')

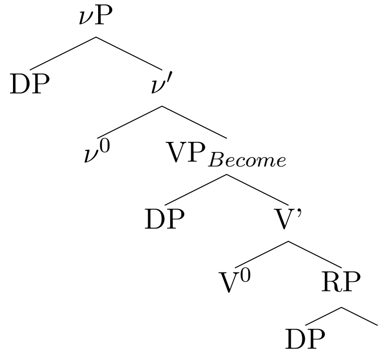
b. **-i/y- class**: *kos-i-ć* ('mow'), *pal-i-ć* ('burn'), *dziw-i-ć* ('amaze')

c. **-ej- class**: *droż-e-ć* ('get expensive'), *głupi-e-ć* ('get stupid'), *grzybi-e-ć* (lit. mushroom-e-inf; 'get senile')

d. **inchoative -n- class**: *marz-ną-ć* ('get frozen'), *głuch-ną-ć* ('get deaf'), *mok-ną-ć* ('get wet')

Following much decompositional work on Argument Structure (cf. e.g. (Ramchand, 2003) and (Marantz, 2003)) we assume a tripartite lower clausal structure as in (2):

(2)



The ν layer, present only in non-unaccusative structures, denotes a causing subevent. VP stands for a transition subevent. RP is a result state where Slavic lexical prefixes and Germanic particles occur (cf. (Ramchand, 2004) and (Svenonius, 2004)). Additional provisos are needed in order to explain the way (inner) subjects are identified and these involve rejecting the Θ-Criterion.

The gist of the hypothesis is that conjugation class suffixes are category defining heads (i.e. verbalizers) in the sense of (Marantz, 1997) that can be merged at different levels in the structure (i.e. either V or ν, see (2)).

(3) a. **high verbalizers**: -aj-, -i/y- → merged as ν^0

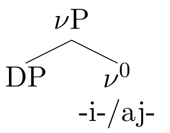
b. **low verbalizers**: -ej-, -n- → merged as V^0

What that means is that roots selecting for high verbalizers will necessarily be bieventive (i.e. the root will name both Causing and Caused subevents). Roots selecting for low verbalizers, on the other hand, will never be able to be augmented upwards and will only name the Caused (Become) predicate. Some roots will be able to take both types of verbalizers. There are several important predictions stemming from this kind of analysis to which we turn directly.

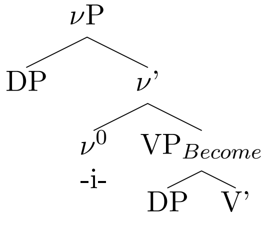
3 Valency of the predicate

There is no straightforward prediction w.r.t. the number of arguments for high verbalizer stems. This is because VP_{Become} subevent is not obligatory. Thus, we could have (i) an unergative monovalent structure in (4) for *bębn-i-ć* ('drum') or *gad-a-ć* ('chat'); (ii) a bievental structure with Subject of Causing event and Subject of Caused event in (5) for *brudz-i-ć* ('dirty'):

(4)



(5)



Yet, there is a prediction for low verbalizer stems to the effect that they can never be transitive (cf. (6a) for inchoative -n- and (6b) for -ej- verbalizer):

(6) a. *Deszcz z-mok-ni-e Jana.
rain$_{NOM}$ pref-wet-n-pres.3sg Jan$_{ACC}$
intended:
'The rain will make Jan wet.'

b. *Jurek o-głupi-ej-e babcię.
Jurek$_{NOM}$ pref-stupid-ej-pres.3sg granny
intended:
'Jurek will stupify granny.'

This allows for a formulation of the following constraint[1]:

(7) $\mathtt{V}_{[low]}$ `*→ DP V DP`

which can be read as: *low verbalizer stems do not enter into structural relations as in (5).*

On the other hand, with roots whose semantics is potentially bieventive and which take both types of verbalizers, we predict that the variant with a high verbalizer will always be transitive, whereas the one with the low one - intransitive:

(8)

Transitive	*gloss*
gas-**i**-ć	put out
o-ślep-**i**-ć	make blind
głodz-**i**-ć	starve
u-piększ-**y**-ć	beautify
Intransitive	*gloss*
gas-**ną**-ć	go out
ślep-**ną**-ć	get blind
głod-ni-**e**-ć	get hungry
pięk-ni-**e**-ć	get beautiful

and the constraints expressing this are in (9), where (9b) is really a subcase of (7):

(9) a. $\mathtt{V}_{[low/high]}$ `*→ DP` $\mathtt{V}_{[high]}$

b. $\mathtt{V}_{[low/high]}$ `*→ DP` $\mathtt{V}_{[low]}$ `DP`

4 The reflexive marker

Another part of the prediction is contingent on specific assumptions about the reflexive marker in Polish. We assume that it is underspecified to the extent that it can be merged in any Subject-of-predicate position, i.e. in Spec, νP as external argument; in Spec,VP in the so-called anticausative/inchoative construction and Reflex-

[1] Henceforth, all constraints are given in typewriter font.

iva Tantum verbs; and in Spec,RP correlating with the presence of a lexical prefix. Additionally, it cen be merged higher than the thematic domain, as a Modal head of sorts, resulting in the so-called Impersonal Reflexive Construction (IRC).

(10) *MoodP - IRC*
$\texttt{V refl DP}_{ACC}$
a. Nastroiło się gitarę.
tune$_{past.3sg.neut}$ refl guitar$_{ACC}$
'Someone/I tuned the guitar.'

(11) *Spec,νP - Middle (Obligatory adverbial)*
$\texttt{DP}_{NOM}\ \texttt{refl Adv V}$
a. Ten chleb się dobrze kroi.
this bread refl well cut$_{pres.3sg}$
'This bread cuts well.'

(12) *Spec, VP - Anticausative*
$\texttt{DP}_{NOM}\ \texttt{V refl}$
a. Gałąź złamała się.
branch break$_{past.3sg.f}$ refl
Spec, VP - Reflexiva Tantum
$\texttt{DP}_{NOM}\ \texttt{V refl}$
b. Jan spocił się.
Jan sweat$_{past.3sg.m}$
'Jan sweated.'

(13) *Spec,RP - prefix-induced*
$\texttt{DP pref-V refl}$
a. Gałąź roze-schła się.
branch pref-dry$_{past.3sg.f}$ refl
'The branch dried into pieces.'
b. Marek za-gadał się.
Marek pref-talk$_{past.3sg.m}$ refl
'Marek forgot himself talking.'

Following the idea in (Manzini, 1986) we assume that the reflexive merged in Spec,νP and higher (in Mood) is [-anaphoric], whereas the anticausative and prefix-induced reflexives are [+anaphoric]. Thus, the predictions are as follows:

- Only high verbalizer stems are able to take a reflexive marker in Spec,νP. This is because low verbalizers simply lack this layer. The lack of this layer is also connected with inability to assign ACC Case.

(14) $\texttt{V}_{[low]}\ {*}{\rightarrow}\ \texttt{DP}_{NOM}\ \texttt{V refl}$

(15) $\texttt{V}_{[low]}\ {*}{\rightarrow}\ \texttt{V DP}_{ACC}$

(16) *Marek zmok-nie-∅ się .
Marek$_{NOM}$ wet-n-pres.3sg refl
intended:
'Someone will wet Marek.'

(17) *Zmok-nie-∅ Marka.
wet-n-pres.3sg Marek$_{ACC}$
intended: 'Marek will get wet.'

- Only high verbalizers are able to take a reflexive marker in Spec,VP. This is due to anaphoric nature of the reflexive, which requires the presence of external argument to bind it.

(18) $\texttt{V}_{[low]}\ {*}{\rightarrow}\ \texttt{DP}_{NOM}\ \texttt{V refl}$

(19) *Gałąź zmok-ni-e się.
branch$_{NOM}$ wet-ni-pres.3sg refl
intended:
'The branch will get wet.'

Furthermore, low verbalizers are predicted never to form Reflexiva Tantum for a similar reason.

- The only two types of reflexive that low verbalizer stems are predicted to be able to take are (i) prefix-induced (in Spec,RP) (cf. (20) and (21)) and (ii) Impersonal reflexive (cf. (22)) with a covert *pro* Subject. Both of the restrictions can be collapsed in (20):

(20) $\texttt{V}_{[low]}\ \rightarrow\ \texttt{DP}_{NOM}\ \texttt{*(pref-)V refl}$

(21) Gałąź roze-sch-ni-e
branch$_{NOM}$ pref-dry-n-pres.3sg
się.
refl
'The branch will dry to pieces.'

(22) Głupi-eje-∅ się.
stupid-ej-pres.3sg refl
'One gets stupid.'

5 Lexical prefixes

There is one property that distinguishes low verbalizer -ej- stems from low verbalizer -n- stems. The former show overt adjectivizing morphemes:

(23) a. *pięk-**ni**-e-ć* (beaut-adj-v-inf; 'get (more) beautiful')
b. *mar-**ni**-e-ć* (miser-adj-v-inf; 'get (more) miserable')

If RP is a part of the functional sequence in the sense that it is selected by VP, and if adjectives do not have the property of selecting RP, it follows that deadjectival -ej- stems will never be able to take RP. Consequently, it follows that they will never be able to take lexical prefixes. Therefore, the only kind of reflexive marker that low verbalizer -ej- stems are allowed to take is the Impersonal one.[2]:

(24) $\texttt{V}_{[-ej-]}$ `*⟶ DP (pref)-V refl`

(25) a. *Ta dziewczyna
this girl
roz-pięk-ni-ej-e się.
pref-beaut-adj-v-pres.3sg refl
intended: 'This girl will get beautiful.'

6 Passives

We assume that both Periphrastic and morphological (Impersonal) Passives in Polish are nominalizers in the sense of (Marantz, 1997), similarly to a deverbal noun in -NIE/CIE (cf. (Jabłońska, in preparation) for arguments). All these three types of nominalizations select for νP. That hypothesis results in the following predictions:

- Periphrastic Passive is only possible with high verbalizer stems. Due to the fact, however, that the lack of Periphrastic Passive might be tied with the necessarily monovalent nature of low verbalizer stems, we do not implement this prediction here. The same reasoning cannot be applied to Impersonal -NO/TO construction.

- Impersonal Passive in -NO/TO is also only possible with high verbalizer stems. This is again, because low verbalizer stems are unaccusative and lack ν layer. This is another incarnation of a well-known fact that passives can only absorb an external argument.

(26) $\texttt{V}_{[low]}$ `*⟶ V-no/to (DP`$_{ACC}$`)`

(27) *O-głupi-a-no /
pref-stupid-ej-no /
*Z-więd-nię-to.
pref-wilt-n-to
intended: 'Someone got stupid / wilted.'

7 Implementation

Large coverage grammar-based parsers are known to have a problem with accuracy. They would often produce a number of parses for a given input but the correct one has to be chosen among the competing structures. In addition, in unification-based systems more than 90% of the parsing time goes to DAG manipulation, yet most of the unifications fail. A lot of CPU time is used for copying constraints, unification, backtracking. Deterministic statistical-based parsers trained on treebanks, on the other hand, would give a single analysis of the input although at the cost of it being the wrong one. What we want to propose in this paper is how one can employ syntactic event decomposition to work in a system without doing a deep grammar processing. Therefore we would like to follow the Marker Hypothesis, which states that all natural languages have a close set of words or morphemes which appear in a limited set of grammatical contexts and which signal this context. We create a parser, whose performance is steered by constraints induced by the verbs' inflectional morphology[3]. As we argue in the previous sections certain verbalizers in Polish restrict the argument structure flexibility.

In order to exemplify this idea we have created a parser that will deal only with the four verb classes. Firstly, however, a system where morphology is involved should have a suitable morphological analyzer. For our purposes we would not manage with an off-the-shelf one, since we are looking for specific morphemes which designate a verb as being a member of a certain conjugation class. Therefore, we have constructed our own verb-class generator. In its present form this is a relational database. For a given verb of a given class it contains all possible forms. It has another function as well, any verb-form (at least of the 4 conjugation classes we work with) can be quickly identified as being a member of the appropriate class. The verb root has also been identified. Thus we have established a surjection, allowing us to label any verb form as belonging to a certain class, see (28):

[2] The only counterexample being *starz-e-ć się* ('get old').

[3] Throughout, the term 'inflectional' is taken to refer to the traditional way of thinking about conjugation classes. For us, especially in the case of deadjectival *-ej-* stems, the distinction between inflection and derivation is no longer valid.

(28) *śpiewał* → **-aj- class**
koszą → **-i/y- class**
głupieliście → **-ej- class**
marzniesz → **-n- class**

and another one-to-many relation from the root to the possible verbalizers it takes (i.e. verb classes it can appear in), as in (8), see (29):

(29) V$_{[high]}$ ← *gas* → V$_{[low]}$

The classes or the verbalizers, as described in sections 2-6, are the ones that impose restrictions of the argument structure of a given verb. These constrains are at the same time the ones that guide our parser. They are relatively few for the 4 classes under scrutiny and we repeat them below as rewrite rules in a context-free grammar:

(30) a. V$_{[low]}$ *→ DP V DP
b. V$_{[low]}$ *→ DP$_{NOM}$ V refl
c. V$_{[low]}$ *→ V DP$_{ACC}$
d. V$_{[low]}$ *→ DP$_{NOM}$ *(pref)-V refl
e. V$_{[-ej-]}$ *→ DP (pref)-V refl
f. V$_{[low]}$ *→ V-no/to (DP$_{ACC}$)
g. V$_{[low/high]}$ *→ DP V$_{[high]}$
h. V$_{[low/high]}$ *→ DP V$_{[high]}$ DP

The constraints in (30) are involved in pruning the search space of the parser while parsing a sentence. In principle, this is not a trivial matter since these refer to deep syntactic structures and abstract away from the surface structure, i.e. the free word order of Polish. Ideally, we would like to be able to write rules concerning the deep structres and leave it to a compiler to create the appropriate for the language surface structure rules. For example, the constraint in (30a), given a context-free grammar with binary rules for a free word order language, will remove the following rules:

(31) a. VP → DP Vbar
b. Vbar → V DP
c. VP → Vbar DP
d. Vbar → DP V

In practice, we have solved this issue by a number of simplifications, since the aim of the paper is not to present a full-fledged parser but rather to illustrate the idea of relying on verb-classes when doing syntactic parsing. Therefore we are looking at only four verb classes, we restrict the word order of the sentences submitted to the system, we avoid using adverbials and large noun phrases, the rule file and lexicon file are relatively small and simple. In other words we deal with a toy system but one that clearly and unambiguously examplifies the underlying idea - verbal morphology can lead to more precise syntactic parsing. A concise view of the system is shown in the Fig. 1 below and the relevant syntactic structures for two sample input sentences are given in (32) and (33).

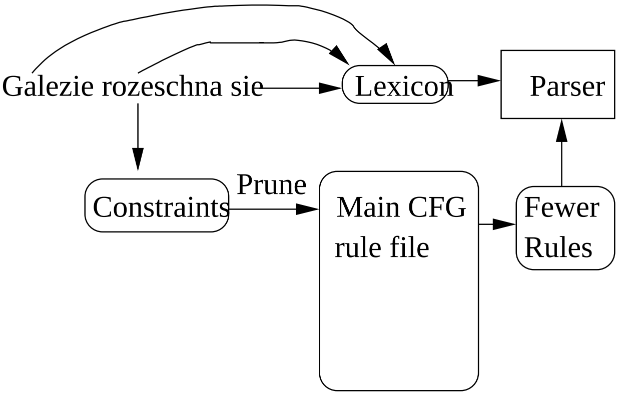
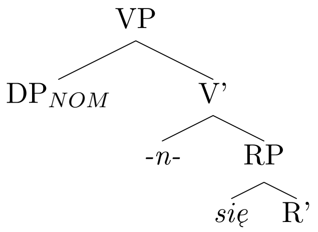
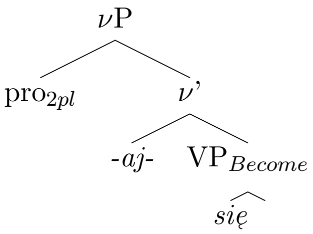


Figure 1: Parser with constraints

(32) Gałęzie rozesch-n-ą się.
branches$_{NOM}$ dry-n-pres.3pl refl
'The branches will get dry.'



(33) Przekon-a-cie się!
convince-aj-pres.2pl refl
'You will get convinced!'



Given a sentence, the system identifies the main verbs in it. These are automatically ascribed a conjugation class, the verb root is also identified. Based on this information, relevant constraints from the constraint store are picked up. The prune module of the system then goes through the large CFG rule file and removes VP rules that do not comply with these constraints. Thus we are left only with the appropriate syntactic configuration in which the verb can be embedded.

The system was implemented in the Oz programming language (van Roy and Haridi, 2004) while for the relational database of the verb forms we used perl. The CFG rules were kept as close as possible to the notational format of the theoretical framework used in sections 2-6 just for our own convenience. The constraints in the constraint store were exchanged with hand-written rules similar to those of the main rule file. The prune module, given a constraint, triggers the *remove-following-rules* function and returns a concise rule file which is then sent to the parser. The lexicon is not exhaustive and diacritics, and other conventions for transcribing Polish were simplified.

8 Conclusion

While there are many details remaining to be worked out, before we can present a full-fledged parser, we think that the idea is worth considering. We are not aware of any previous work where inflectional morphology is taken into consideration while doing syntactic parsing. At the same time we rely on a theoretical framework which has not been favored by the NLP community (except the work of (Stabler, 1997)). While we are not implementing a Principles and Parameters parser, we borrow such ideas from this framework, which we consider very useful and well-grounded. From a computational point of view, this approach is valuable in two ways. Firstly, the verbal entry in the lexicon is considerably impoverished, the specification being restricted to one feature in essense, i.e. [+/-high]. Secondly, the combinatorial explosion of the grammar rules is handled by the constraints coming from the verb classes. Since Polish exibits a relatively free word-order, the context-free rules for the parser will be much more than if we deal with English, for example. In this way inflectional morphology that is traditionally considered an unnecessary burden for the language learner is conceived of as providing significant clues to prune a relatively large search space. We reduce the dimension of the lexicon and achieve a correct syntactic analysis which derives the proper semantic interpretation of the arguments involved, as well as allows the root to remain considerably underspecified semantically in the lexicon.

References

Maya Arad. 1998. *VP structure and the syntax-lexicon interface.* Ph.D. thesis, University College London.

Hagit Borer. 2003. Structuring Sense. An Exo-Skeletal Trilogy. Ms., USC.

Patrycja Jabłońska. in preparation. Causatives and event-decomposition in syntax. Ms., University of Tromsø.

Beth Levin and Malka Rappaport. 1995. *Unaccusativity: At the syntax-lexical semantics interface.* MIT Press, Cambridge, MA.

Maria Rita Manzini. 1986. On Italian *Si.* In Hagit Borer, editor, *The syntax of pronominal clitics*, volume 19 of *Syntax and Semantics*, pages 241–262. Academic Press, Orlando, Florida.

Alec Marantz. 1997. No Escape from Syntax: Don't try morphological analysis in the privacy of your own lexicon. *Penn Working Papers in Linguistics*, 4(2):201–225.

Alec Marantz. 2003. Argument structure. talk presented at Forskerutdanningsseminar, University of Tromsø, May 28.

Gillian Ramchand. 2003. First Phase Syntax. ms, University of Oxford.

Gillian Ramchand. 2004. Time and the event: The semantics of Russian prefixes. to appear in Nordlyd: www.ub.uit.no/munin/nordlyd.

Ed Stabler. 1997. Derivational Minimalism. In Retoré, editor, *Logical Aspects of Computational Linguistics*, pages 68–95. Springer.

Peter Svenonius. 2004. Slavic prefixes inside and outside vp. to appear in Nordlyd: www.ub.uit.no/munin/nordlyd.

Peter van Roy and Seif Haridi. 2004. *Concepts, Techniques and Models of Computer Programming.* MIT Press.

Syntax-Semantics Interface in Lexicalist Theories

Valia Kordoni
Department of Computational Linguistics
Saarland University
P.O. Box 15 11 50
D-66041 Saarbrücken
Germany
kordoni@coli.uni-sb.de

Abstract

This paper focuses on valence alternations in German, Modern Greek, and English.

1 Introduction

This paper focuses on valence alternations in German (examples (1)-(4)), Modern Greek (examples (5)-(6)), and English (examples (7)-(8)).

(1) Peter goß die Blumen mit Wasser.
Peter poured the flowers.A with water
"Peter watered the flowers".

(2) Peter goß Wasser auf die Blumen.

(3) Peter füllte den Tank (mit Wasser).
Peter.N filled the tank.A (with water)
"Peter filled the tank (with water)".

(4) Peter füllte Wasser in den Tank.

(5) O georgos fortose to ahiro
the farmer.N load.PAST.3S the hay.A
sto karo.
onto-the wagon
"The farmer loaded the hay on the wagon".

(6) O georgos fortose to karo
the farmer.N load.PAST.3S the wagon.A
me ahiro.
with hay
"The farmer loaded the wagon with hay".

(7) The farmer loaded the wagon with hay. (*with*-variant)
⇒ The hay was loaded on the wagon.
⇒ The wagon was loaded with hay.

(8) The farmer loaded hay on the wagon. (locative variant)
⇒ The hay was loaded on the wagon.
⇏ The wagon was loaded with hay.

These alternations involve direct internal (i.e., objects), as well as indirect prepositional arguments: NP_k V NP_i [P NP_j] → NP_k V NP_j [P NP_i].[1] Such alternation patterns in German, Modern Greek and English characterize among others the behaviour of verbal predicates which participate in the so-called Locative Alternation phenomena[2] (see Dowty (1991), Rappaport and Levin (1988), Levin and Rappaport Hovav (1991)). For instance, alternations in German with the locative verbs *gießen* (pour/water) and *füllen* (fill) are of the general form presented above. Two main features of these verbs in German, – as well as in English, and Modern Greek –, are that they are morphologically identical and that they involve two arguments: one denoting a *location* and one denoting the *locatum* (*die Blumen* (flowers)/*den Tank* (tank) and *Wasser* (water), respectively, in (1)-(4) above).

Valence alternations like the ones in (1)-(8) have always posed an interesting theoretical challenge. As Rappaport and Levin (1988) have shown, the locative alternation variants in English differ in entailments: the *with* variant has an entailment the locative alternant lacks (see examples (7) and (8), respectively). Based on this, Rappaport and Levin (1988) and Pinker (1989) assume that the two alternants of the English locative verbs *load* and *spray* have different semantic contents and propose that the alternation is about alternate choices of object (see examples (9) and (10)). The problem, though, with such analyses of valence alternations is that there is no independent semantic motivation for the new metalanguage predicate/keyword BY (see (9) below).

(9) Peter sprayed the statue with paint. (*with*-variant)
ACT-ON (PETER, STATUE, BY (CAUSE (PE-

[1] The indices denote referential identity.

[2] As well as in the Dative Shift phenomena, which we do not examine here due to lack of space.

TER, GO (PAINT, TO (STATUE)))))

(10) Peter sprayed the paint onto the statue. (locative variant)
CAUSE (PETER, GO (PAINT, TO (STATUE)))

2 Valence Alternations in Lexical Mapping Theory (LMT)

In the Lexical Mapping Theory (LMT) literature (English) locative alternations are not extensively discussed. In an effort to account for such alternations in LMT of LFG, adapting the thematic role analysis which Ackerman (1992) has proposed for locative inversion in Hungarian to the locative alternation data at hand is a natural step to take and gives results along the lines described in (11) and (12). As shown in (11) and (12), though, such a thematic role analysis is indeed problematic because the attempt to account for two different linkings to the respective grammatical functions from the same array of thematic roles clearly fails.

(11) The farmer loaded the wagon with hay. (*with*-variant)

load<	agent	theme(locatum)	location >
	-o	??	??
	SUBJ	OBL$_{with/theme}$	OBJ

(12) The farmer loaded the hay on the wagon. (locative variant)

load<	agent	theme(locatum)	location >
	-o	-r	-o
	SUBJ	OBJ	OBL$_{on/goal}$

Our first aim here is to address the problematic points that a traditional LMT account does not seem able to avoid. To do that: (i) we rely on Rappaport and Levin's (1988) conclusion that the locative alternation variants differ in entailments, as well as on the fact that this difference in entailments is found across all locative alternation verbs in English (this is also true for Modern Greek and German); this is a fact which according to Rappaport and Levin (1988) suggests that the entailments in the case of locative alternation verbs are associated with the variants and not the verbs or the different arguments these verbs support, as is for instance the case with the dative alternation in English; (ii) we follow Baker (1997), Maling (2001), and Levin and Rappaport Hovav (2001) who suggest that with locative alternation verbs either the location or locatum argument shows "object" properties depending on which is object (see examples (13) and (14) which are due to Baker (1997) and Williams (1980); their counterparts in German, for instance, we give in (15) and (16)).

(13) the loading of hay onto wagons/the loading of wagons with hay (nominalization)

(14) John loaded the hay onto the wagon green./John loaded the wagon full with hay. (secondary predication; from Williams (1980))

(15) das Laden von Heu auf den Wagen/das Beladen des Wagens mit Heu

(16) Peter lud den Wagen mit Heu voll.

The LMT analysis proposed in (17) and (18) adopts these two points. Moreover, following Zaenen (1993), the proposal for both variants of the German locative verb *gießen*, for instance, does not rely on thematic roles. Instead, conventional labels in the spirit of Zaenen (1993), such as *agent*, *patient* and *nonpatient*, are used in order to indicate that the verb supports three arguments, each of which is associated with some general lexico-semantic entailments: an agent ("external"/ "semantically-and-syntactically-most-prominent" argument (a $\hat{\theta}$ [-o] argument in LMT terms)), and two other arguments, one with patient entailments (*patient*), and one with neither patient nor secondary-patient entailments (*nonpatient*). Consequently, *nonpatient* is correctly predicted in both cases to bear the intrinsic classification feature [-o], which maps it to the grammatical function OBL in the case of both variants of the German locative verb *gießen*. *patient*, on the other hand, which can be related either to the argument of the verb which denotes the locatum (see (17)) or to the argument of the verb which denotes the location (see (18)), since both may bear "object" properties, when they are not instantiated as indirect prepositional arguments, as we have seen above, is intrinsically classified as [-r]. This classification maps it to the grammatical function OBJ in the case of both variants of the German locative verb *gießen*. This treatment is in accordance with the proposal of Baker (1997), Maling (2001), and Levin and Rappaport Hovav (2001) for this argument of locative alternation verbs which we presented above briefly.

(17) Peter goß Wasser auf die Blumen. (locative alternant)

gießen<agent	patient (locatum)	nonpatient> (location)
-o ($\hat{\theta}$-arg)	-r	-o
SUBJ	OBJ	OBL$_{(auf)}$

(18) Peter goß die Blumen mit Wasser. (*mit* (with)-variant)

	gießen<agent	patient (location)	nonpatient> (location)
	-o ($\hat{\theta}$-arg)	-r	-o
	SUBJ	OBJ	$\mathrm{OBL}_{(mit)}$

The analysis proposed in (17) and (18) addresses the problems that traditional LMT accounts have encountered (see (11) and (12)), since it can account for the two different linkings of locative alternation verbs to their respective grammatical functions from the same array of lexico-semantic entailments and intrinsic classification features for all the arguments of the locative verbs.

It does not have, though, much to say about the correct intuitions of proposals such as that of Rappaport and Levin (1988), according to which the difference in the entailments associated to the locative alternation verbs is related to the variants as whole constructions and not to their verbal heads or the different arguments these verbal heads support, as is for instance the case with the dative alternation in English. And this is a general and genuine problem related to the Lexical Mapping Theory (LMT) of LFG. In LMT the properties of grammatical constructions are (almost) impossible to be captured and accounted for in a constrained way.

3 Valence Alternations in HPSG and MRS

In the rest we show that the theoretical framework of HPSG (Pollard and Sag (1994)), instead, with semantic representations in Minimal Recursion Semantics (MRS; Copestake et al. (1999)) constitutes the appropriate theoretical basis for a robust, linguistically-motivated account of locative alternations cross-linguistically (see examples (1)-(8)), which does not only overcome the natural limitations of other syntactic and semantic analyses of such constructions (see among others Rappaport and Levin (1988), Pinker (1989), Markantonatou and Sadler (1996)), but also provides the necessary formal generalizations for the analysis of constructions in a multilingual context, since MRS structures are easily comparable across languages.

Thus, the account we propose here for locative alternations in German, Modern Greek, and English (see examples (1)-(8) above) follows the proposal of Koenig and Davis (2000) for valence alternations. Their analysis is based on a minimal recursion approach to semantic representation and is formalized using the Minimal Recursion Semantics (MRS) framework of Copestake et al. (1999) (see also Copestake et al. (2001)). In brief, Minimal Recursion Semantics is a framework for computational semantics, in which the meaning of expressions is represented as a flat bag of Elementary Predications (or EPs) encoded as values of a LISZT[3] attribute. The denotation of this bag is equivalent to the logical conjunction of its members. Scope relations between EPs are represented as explicit relations among EPs. Such scope relations can also be underspecified. The assumption of current MRS is that each lexical item (other than those with empty EP bags) has a single distinguished main EP, which is referred to as the *KEY* EP. All other EPs either share a label with the KEY EP or are equal to some scopal argument of the KEY EP.

According to Koenig and Davis, for situation-denoting EPs, which are also most interesting for our purposes here, the following generalizations hold: (i) EPs do not encode recursively embedded state-of-affairs (SOAs); (ii) EPs can have one, two, or three arguments; (iii) if an EP has three arguments, then one of them is a state-of-affairs, and another is an undergoer co-indexed with an argument of the embedded state-of-affairs. Finally, as far as direct arguments are concerned, in Koenig and Davis (2000) these are predicted to link off the value of the KEY attribute.

Following the *lexical list hypothesis* of Koenig and Davis (2000), according to which lexical items include more than one EPs in their semantic content, but lexically they select only one of these EPs as their KEY, we propose that the semantic properties of the arguments of the verb *gießen* (water) in example (1), for instance, are captured by the following semantic type:

(19) CONTENT value of *gießen*

$$\begin{bmatrix} \text{KEY } \boxed{3} \begin{bmatrix} \textit{gießen-ch-of-st-rel} \\ \text{ACT } \boxed{1} \ (\textit{Peter}) \\ \text{UND } \boxed{2} \ (\textit{die Blumen}) \end{bmatrix} \\ \text{LISZT } \left\langle \boxed{3}, \begin{bmatrix} \textit{mit-rel} \\ \text{ACT } \boxed{1} \\ \text{UND } \boxed{4} \\ \text{SOA } \boxed{3} \end{bmatrix}, \begin{bmatrix} \textit{gießen-ch-of-loc-rel} \\ \text{ACT } \boxed{1} \\ \text{FIG } \boxed{4} \ (\textit{Wasser}) \end{bmatrix} \right\rangle \end{bmatrix}$$

(19) above captures that the alternant of the German locative verb *gießen* whose indirect internal argument is headed by the preposition

[3] RELS nowadays.

mit (with) and whose direct internal argument is a *location* (example (1)) denotes situations that must be both changes of state and changes of location.

The locative alternant of the verb *gießen* (example (2)), on the other hand, denotes only a simple change of location. This is captured by the semantic type in (20) below.

(20) CONTENT value of *gießen* (locative variant)[4]

$$\begin{bmatrix} \text{KEY } \boxed{5} \begin{bmatrix} \textit{gießen-ch-of-loc-rel} \\ \text{ACT } \boxed{1}\ (\textit{Peter}) \\ \text{FIG } \boxed{4}\ (\textit{Wasser}) \end{bmatrix} \\ \text{LISZT } \langle \boxed{5} \rangle \end{bmatrix}$$

The analysis presented above holds also for both alternants of the verb *füllen* (fill; examples (3) and (4)), as well as for the Modern Greek and English data in examples (5)-(8). For the *mit* (with) alternant of the verb *füllen* (example (3)), where the indirect internal argument (the PP *mit Wasser*) appears to be optional, we assume that semantically the *ch-of-loc* EP carries existential import, even when the PP is not syntactically overt.

The analysis we have presented above accounts also for other subclasses of locative alternating verbs, among them the so-called removal predicates (e.g., predicates like *wischen*, *säubern* in German, *wipe* in English, *skupizo* in Modern Greek), as well as the so-called impingement predicates (e.g., the predicates *schlagen* in German, *hit* in English, *htipo* in Modern Greek). Here we concentrate on the former class.

(21) *Peter wiped the pan of the grease.

(22) Peter wiped the grease from the pan.

(23) *Peter wischte die Tafel von Kreide.
Peter.N wiped the board.A from chalk
"*Peter wiped the board of chalk".

(24) Peter wischte die Kreide von der
Peter.N wiped the chalk.A from the
Tafel.
board
"Peter wiped the chalk from the board".

(25) *O Petros skupise to tigani apo
the Peter.N wipe.PAST.3S the pan.A from
to ladi.
the oil
"*Peter wiped the pan of the oil".

(26) O Petros skupise to ladi apo
the Peter.N wipe.PAST.3S the oil.A from
to tigani.
the pan
"Peter wiped the oil from the pan".

The predicates in (21)-(26) above denote a contact with the *location*, as well as a change of location. These predicates may also specify the manner or the instrument related to the action of moving. For instance, the English removal predicate *wipe*, the German *wischen* (wipe), as well as the Modern Greek *skupizo* (wipe), do not admit an indirect argument (*of/von*-PP complement) when their *location* argument is realized as their direct internal argument (object; examples (21), (23), (25)). In this case *wipe*, as well as *wischen* and *skupizo*, do *not* entail the existence of a *locatum* argument. For instance, the act of wiping a board does not necessarily result in wiping something off it.

However, the removal predicates *trim* (in English), *säubern* (in German), and *katharizo* (in Modern Greek) are different than *wipe*, *wischen*, and *skupizo*, respectively, in the sense that "trimming an object" necessarily means "trimming something off this object":

(27) Peter trimmed the bush of the dry branches.

(28) Peter säuberte den Busch von
Peter.N trimmed the bush.A of
trockenen Ästen.
dry branches
"Peter trimmed the bush of dry branches".

(29) O Petros katharise to thamno
the Peter.N trim.PAST.3S the bush.A
apo ta xera kladia.
of the dry branches
"Peter trimmed the bush of the dry branches".

Thus, we propose that the semantic properties of the arguments of the verbs *wipe*, *wischen* and *skupizo*, which denote a change of location, when a *locatum* argument is realized as their direct internal argument, can be captured by a type like the following:

(30) CONTENT value of *wipe*

$$\begin{bmatrix} \text{KEY } \boxed{5} \begin{bmatrix} \textit{wipe-ch-of-loc-rel} \\ \text{ACT } \boxed{1}\ (\textit{Peter}) \\ \text{FIG } \boxed{4}\ (\textit{the grease}) \end{bmatrix} \\ \text{LISZT } \langle \boxed{5} \rangle \end{bmatrix}$$

säubern (trim; see example (28) is different than *wischen*:

[4] FIG(ure) in a *ch(ange)-of-loc(ation)-rel(ation)* denotes the argument which is moving changing location.

(31) CONTENT value of *säubern*

$$\begin{bmatrix} \text{KEY } \boxed{3} \begin{bmatrix} \textit{säubern-ch-of-st-rel} \\ \text{ACT } \boxed{1} \, (\textit{Peter}) \\ \text{UND } \boxed{2} \, (\textit{den Busch}) \end{bmatrix} \\ \text{LISZT } \langle \boxed{3}, \begin{bmatrix} \textit{von-rel} \\ \text{ACT } \boxed{1} \\ \text{UND } \boxed{4} \\ \text{SOA } \boxed{3} \end{bmatrix}, \begin{bmatrix} \textit{säubern-ch-of-loc-rel} \\ \text{ACT } \boxed{1} \\ \text{FIG } \boxed{4} \, (\textit{Ästen}) \end{bmatrix} \rangle \end{bmatrix}$$

That is, as (31) above captures, in German trimming necessarily results in trimming something off something else; in the case of example (28) above trimming the bush results in trimming the dry branches off the bush. And this is what the semantic type in (31) captures. The semantic properties of the English verb *trim* and the Modern Greek verb *katharizo* can also be captured by a type like the one in (31) adapted to English and Modern Greek, respectively.

4 Conclusions

We have shown that the theoretical framework of HPSG (Pollard and Sag (1994)) enriched with semantic representations in Minimal Recursion Semantics (MRS; Copestake et al. (1999), Copestake et al. (2001)) constitutes the appropriate theoretical basis for a robust, linguistically-motivated account of locative alternations, which provides the necessary formal generalizations for the analysis of (such) constructions in a multilingual context, since MRS structures are easily comparable across languages. To show this we have considered *contact* and *removal* constructions in German, Modern Greek, and English (examples (1)-(8) and (21)-(29)).

As a general comment we need to underline that the MRS-based analysis we have presented above allows for a linguistically-motivated account of the syntactic properties of apparent semantic doublets, which avoids the processing load problems that are inseparable from (directional or even bi-directional à la Flickinger (1987)) lexical rule approaches to parsing constructions containing indirect arguments in particular and to development of (the lexicon of) large-scale (computational) grammars of natural language based on HPSG in general. Consequently, (the lexicon of) large-scale (computational) grammars may become more efficient, since it needs to depend on fewer or even no lexical rules at all, and thus less complicated for the grammar writer to maintain, as well as to develop further. Here we focussed on (some of) the theoretical assumptions upon which the achievement of such a goal can be based realistically.

Finally, we have shown that HPSG enriched with semantic representations in MRS can capture the properties of grammatical constructions and account for them in a natural way. That is, in the MRS-enriched HPSG analysis that we have sketched above:

1. the implicational differences of the alternations are derived from alternative realizations, not from alternative lexical meanings;
2. both the location and the locatum arguments are shown to bear "object" properties depending on which is object; that is, the insights of Rappaport and Levin (1988), Baker (1997), Maling (2001), Levin and Rappaport Hovav (2001) seem to be correct;
3. grammatical constructions and their semantics are treated in a constrained way.

The analysis we have presented above shows clearly that the semantic and linking models in LFG and HPSG vary not only ontologically, but also in the range of linguistic phenomena they attempt to explain.

References

Farell Ackerman. 1992. Complex predicates and morphological relatedness: Locative alternation in hungarian. In Ivan A. Sag and Anna Szabolcsi, editors, *Lexical Matters. CSLI Lecture Notes no. 24*, pages 55–84. CSLI Publications, Stanford, Calif.

Mark Baker. 1997. Thematic roles and syntactic structures. In L.Haegeman, editor, *Elements of Grammar. Handbook of Generative Syntax*, pages 73–137. Kluwer, Dordrecht.

Ann Copestake, Dan Flickinger, Ivan A. Sag, and Carl J. Pollard. 1999. Minimal Recursion Semantics: An Introduction. Ms., Stanford University.

Ann Copestake, Alex Lascarides, and Dan Flickinger. 2001. An algebra for semantic construction in constraint-based gram mars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001)*, pages 252–259. Toulouse, France.

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67:547–619.

Daniel Flickinger. 1987. *Lexical Rules in the Hierarchical Lexicon.* Ph.D. thesis, Stanford University, California.

Jean-Pierre Koenig and Anthony R. Davis. 2000. The KEY to Lexical Semantics. Paper presented at the 7th International Conference on Head-Driven Phrase Structure Grammar, held on July 22-23, 2000 as part of the Berkeley Formal Grammar Conference 2000.

Beth Levin and Malka Rappaport Hovav. 1991. Wiping the Slate Clean: A Lexical Semantic Exploration. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics*, pages 123–152. Blackwell, Cambridge MA and Oxford UK.

Beth Levin and Malka Rappaport Hovav. 2001. What Alternates in the Dative Alternation? Ms., Colloquium Series, Department of Linguistics and Philosophy, MIT, Cambridge, MA, November 9, 2001.

J. Maling. 2001. Dative: The Heterogeneity of the Mapping Among Morphological Case, Grammatical Functions, and Thematic Roles. *Lingua*, 111:419–464.

S. Markantonatou and L. Sadler. 1996. Linking Indirect Arguments. *Essex Research Reports in Linguistics*, 9:24–63.

Steven Pinker. 1989. *Learnability and Cognition: the acquisition of argument structure.* Cambridge, MA: MIT Press.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago.

Malka Rappaport and Beth Levin. 1988. What to do with θ-roles. In Wendy Wilkins, editor, *Thematic Relations. Syntax and Semantics 21*, pages 7–36. Academic Press Inc.

Edwin Williams. 1980. Predication. *Linguistic Inquiry*, 11:203–238.

Annie Zaenen. 1993. Unaccusativity in dutch: Integrating syntax and lexical semantics. In James Pustejovsky, editor, *Semantics and the Lexicon*, pages 129–162. Kluwer Academic Publishers, Dordrecht.

Automatic Extraction of Subcategorization Frames from Spoken Corpora

Jianguo Li
Department of Linguistics
The Ohio State University
1712 Neil Ave
Columbus, OH, 43210
jianguo@ling.ohio-state.edu

Chris Brew
Department of Linguistics
The Ohio State University
1712 Neil Ave
Columbus, OH, 43210
cbrew@acm.org

Abstract

We built a system for automatically extracting subcategorization frames (SCFs) from corpora of spoken language. The acquisition system, based on the design proposed by Briscoe & Carroll (1997) consists of a statistical parser, a SCF extractor, an English lemmatizer, and a SCF evaluator. These four components are applied in sequence to retrieve SCFs associated with each verb predicate in the corpus. An initial experiment, based on a sample of 14 verb types and 330 verb tokens, tests the hypothesis that this will work acceptably for spoken language. The current prototype system achieves accuracy (type precision and recall and token recall) comparable to previous systems that work with written corpora.

1 Introduction

SCFs, specifying the number and type of arguments that a particular predicate requires, are considered a key component of computational lexicon and have benefited various natural language processing (NLP) tasks.

SCFs are essential for the development of parsing technology. Manning (1993) showed that access to SCFs helps rule-based parsers to constrain the number of analyses that are generated. For example, a rule-based parser, based solely on SCFs, can deduce that the PP *on the table* in *Mary put the box on the table* is a complement of the verb *put*, not a noun phrase modifier. Briscoe & Carroll (1997) showed that the SCF frequency information helps improve the accuracy of their statistical parser.

In addition to parsing, SCFs can also benefit other domains of NLP. For example, Lapata & Brew (2004) used SCFs in their verb sense disambiguation task. They built a model to disambiguate verb senses by exploring the association between the syntactic frames that a particular verb licenses and the semantic classes it belongs to. In case of the verb *serve*, we can guess the semantic class (Levin 1993) solely on the basis of the syntactic frames with which they appear with. As shown in the examples below, serve in (1a) appears with [NP NP] and belongs to GIVE verbs, in (1b) it occurs with [NP] and is a member of FIT verbs, in (1c) it appears with [NP PP(as)] and is a MASQUEADE verb and finally it occurs with [NP PP(to)] in (1d) and belongs to the FULFILLING verbs.

(1a) They serve our guest an Italian meal.
(1b) The airline serves 164 destinations.
(1c) He served Napoleon as Minister of Interior.
(1d) He served an apprentice to a photographer.

Several conventional syntax dictionaries exist for English. Nevertheless, there is still a need for a program that can automatically extracts SCFs from corpora because SCFs dictionaries do not always reflect actual language use. First, manually constructed syntax dictionaries usually do not exhaust all the SCF possibilities. For example, none of the syntax dictionaries lists [S(*what*)] for the verb *add*. However, it is a valid SCF for *add* in a sentence like *he adds what he thinks is right* (Korhonen 2002). Next, no manually constructed syntax dictionary encodes the relative frequency of each listed SCF. *Oxford Advanced Learner's Dictionary* (OALD) lists both *agree about* and *disagree about*, but *agree about* never occurs in the learning corpus (Manning 1993). This observation led Manning to conclude that people like to *agree with* people, but *disagree about* topics. These problems suggest that automatic acquisition of SCFs is necessary as a complement to intuition.

The current system uses a spoken corpus. An SCF dictionary built from spoken language may, if of acceptable quality, be of particular value in NLP tasks involving syntactic analysis of spoken language. Since the statistical parser (Charniak's) that we use was not designed for spoken language, we wished to test the hypothesis that the system as a whole will work for the different demands of spoken language.

2 Description of the system

Methods for automatic acquisition of SCFs usually proceed in the following two steps: Methods for automatic acquisition of SCFs usually proceed in the following two steps:

I. Hypothesis Generation: Identify all SCF cues for a particular verb.
II. Hypothesis Selection: Determine which SCF cue is a valid SCF for a particular verb.

2.1 Overview

The current SCF acquisition system consists of four components that are applied in sequence to the spoken language of BNC to retrieve SCFs associated with each verb predicate in the corpus:

- **A Statistical Parser**: Charniak's Parser, trained on tree-bank, returns a parse for each sentence. The output does not explicitly distinguish between complements and adjuncts (Charniak 1997).
- **A SCF Extractor**: An extractor is used to identify each verb predicate in parsed trees and identify the syntactic category for all its sisters and combine them into a SCF cue.
- **A Lemmatizer**: MORPHA, an English morphological analyzer, is used to return lemma for each verb (Minnen, Carroll, Pearce 2001).
- **A SCF Evaluator**: An evaluator is used to filter SCF cues on the basis of their reliability and likelihood.

2.2 Hypothesis generation

The hypothesis generation consists of the first three components: Charniak Parser, SCF Extractor and Morpha. It takes raw corpus data as its input and generates SCF cues as its output. For example, building SCF entries for *give* and given that one of the sentences in our training data is (2a), the parser first returns a parsed tree for this sentence, as shown in (2b). The extractor then builds a SCF cue in (2c) based on the parsed tree. After the lemmatizer replaces the verb *gave* with its lemma *give*, the SCF cue appears as in (2d), which serves as the input to the hypothesis selection.

(2a) Brenda gave us a talk.
(2b) (S (NP (NNP Brenda))
        (VP (VB gave)
            (NP (PRP us)
            (NP (DT a)
                (NN talk))))
(2c) gave: [NP NP]
(2d) give: [NP NP]

2.2.1 Extractor

The SCF extractor in this task merits further explanation. In some cases, the extractor has to do some extra work in order to build SCF cues we are seeking for.

- **Finite and Infinite Clauses**: Charniak parser uses **S** and **SBAR** to label different type of clauses, which obscures too many details of the internal structure of each clause. The Extractor is thereby modified to identify the internal structure of each constituent labeled as **S** or **SBAR**.

Parser's Label	**Clause Type**	**Clause Subtype**		**Desired SCF**
S	Infinite Clause	Control		[(NP) INF(*to*)]
		Gerundive clause		[(NP) V-VING]
		Small clause		[(NP) ADJP|PassP]
SBAR		Control		[(NP) INF(*wh-to*)]
	Finite Clause	Complementizer	*wh*	[S(*wh*)]
			that	[S(*that*)]
		No complementizer		

Table 1: SCFs for different clauses

Example: asked me to say …
Parsed Tree: (VP (VBN asked)
        (S (NP (PRP me))
            (VP (TO to)
                (VP (VB say) …
SCF cues: ask: [S] → ask: [NP INF(to)]

- **Passive Sentences**: Passive sentences are not indicated in the parsed trees. If the Extractor fails to discover passive structures, the SCF cues associated with passive structures would always miss one complement. To restore the missing complement, the Extractor first locates all verbs tagged as VBN and VBD. It then searches for the nearest preceding auxiliary verb and checks if it is some form of *be*. If so, the Extractor adds an NP to the SCF cue.
Example: He had been saved.
Parsed Tree: (S (NP (PRP He))
        (VP (AUX had)
            (VP (AUX been)
                (VP (VBN saved))))
SCF cues: save: [] → save: [NP]

- **Auxiliary-like Verbs**: Verbs such as *going*, *got*, *used*, when followed by an infinitive clause headed by *to*, act as auxiliary verbs. These SCF cues are simply ignored in this task.
Example: … going to record our meeting.
Parsed Tree:

```
(VP (VBG going)
    (S (VP (TO to)
           (VP (VBG record)
               (NP (PRP our)
                   (NN meeting)))))
```

SCF cues: go: [S] → go: [INF(*to*)] → NONE

- **Verbal Conjunction**: Two or more verbs in a verbal conjunction sometimes share the same complements. However, such information is missing in the parser's output. The Extractor is made to identify the shared complements and then associate them with each individual verb.
  Example: His father bought and sold cars.
  Parsed Tree:

```
(S (NP (DT his)
       (NN father))
   (VP (VBD bought)
       (CC and)
       (VBD sold)
       (NP (NNS cars))))
```

SCF cues: buy: [CC VBD NP]
→ buy/sell: [NP]

- **Phrasal Verbs**: Many English verbs take a particle or an adverb and together they form a phrasal verb. In this task, phrasal verbs are treated as ordinary verbs. The Extractor first extracts the lexical head of all constituents labeled as PRT or ADVP and combines them with the lexical verbs into phrasal verbs. If the phrasal verb is listed in a phrasal verb dictionary (*Cambridge Phrasal Verb Dictionary*), the Extractor builds a SCF cue for the phrasal verb.
  Example: … as she goes along.
  Parsed Tree:

```
(SBAR (IN as)
      (S (NP (PRP she))
         (VP (VBZ goes)
             (ADVP (RB along)))))
```

SCF cues: go: [ADVP(*along*)] → go-along: []

2.3 Hypothesis selection

The hypothesis selection consists of only the SCF Evaluator. A SCF cue generated by the Extractor may be a correct SCF, or it may contain some adjuncts, or it may simply be wrong due to errors made by the parser. Given that Charniak parser makes no distinction between complements and adjuncts and the current system works with only spoken language, SCF cues proposed by the Extractor are likely to contain more noise. Thus, the Evaluator must be able to distinguish between complements and adjuncts, as well as filter out false SCF cues. The Evaluator for the current system is made up of two parts: the BHT (Brent 1992) and a back-off algorithm (Sarkar & Zeman 2000).

- **Binomial Hypothesis Test (BHT)**: Let *p* be the probability that a scf_i will occur with a verb which is not supposed to take scf_i. If a verb occurs *n* times and *m* of those times it co-occurs with scf_i. The probability that all the scf_i cues are false cues is bounded by the binomial distribution:

$$P(m+, n, p) = \sum_{k=m}^{n} \frac{n!}{m!(n-m)!} p^{m}(1-p)^{n-m}$$

If the value of $P(m+, n, p)$ is less than or equal to some small threshold value, then the null hypothesis that verb_j does not take scf_i is extremely unlikely to be true. Hence, scf_i must be a valid SCF for verb_j. Setting a threshold of less than or equal to 0.05 yields a 95% or better confidence to believe that verb_j has been observed frequently enough with scf_i for it to be a valid SCF for verb_j.

The first two figures *m* and *n* can be straightforwardly calculated from the output of the Extractor, but the value of *p* is not that easy to come by. Following Manning (1993), we empirically determined the value of *p* for each SCF cue.

- **Back-off Algorithm**: The SCF cues generated by the Extractor always contain some adjuncts. However, for each such SCF cue, one of its subsets is most likely to be the correct SCF we are seeking for. Table 2 gives some SCF cues generated for the verb *introduce*:

SCF cues proposed by the Extractor	**Correct SCF**
[NP PP(*from*)]	[NP]
[NP PP(*at*) S(*before*)]	
[NP S(*when*)]	
[NP PP(*to*) PP(*in*)]	[NP PP(to)]
[NP PP(*to*) PP(*before*)]	
[NP PP(*to*) ADVP]	

Table 2: SCF cues and correct SCFs for *introduce*

In terms of the implementation of this back-off algorithm, for each verb, we first consider the longest SCF cue proposed by the Extractor. Let assume that this SCF cue fails the BHT. We then eliminate the last constituent from the rejected SCF cue and transfer the frequency of the rejected SCF cue to its successor and submit the chosen successor to the BHT again. In this way, frequencies are most likely to accumulate and valid frames to survive the BHT.

An important issue in the back-off algorithm is the selection of the successor for the rejected SCF cues. In English, word order is relatively rigid and complements tend to appear before adjuncts in a SCF cue. For this reason, we choose always eliminate the last constituent in a SCF cue and submit the resulting successor to the BHT. However, there are a few cases where adjuncts are found to be closer to head verbs. For example, in English, adverbs often have the option of occurring in several different positions within a sentence. Since verbs rarely take adverbs as complements, during hypothesis generation, the Extractor moves all ADVPs to the end of each SCF cue. In doing so, ADVPs always get eliminated first if the proposed SCF cue is rejected.

3 Results and discussion

To evaluate the methods above, we used 1 million words of training data taken out of the spoken corpus of BNC. In this training set, there are 109, 116 verb tokens with 4, 134 verb types. Among them, 907 verb types are seen 10 or more times. The acquisition system acquired 1, 797 SCFs for the 907 verb types (an average of 1.98 per verb).

Table 4 shows the number of *true positives* (TPs), the correct SCF types proposed by out systems, *false positives* (FPs), the incorrect SCF types proposed by our system and *false negatives* (FNs), the correct SCF types not proposed by our system. To calculate type precision and type recall, Briscoe & Carroll (1997) randomly selected 14 verbs with multiple SCF types. We evaluated the results of these 14 verbs against the SCF entries for these verbs in COMLEX (Grishman, Macleod, Meyers, 1994) syntax dictionary. The results are summarized in table 3:

With SCF acquisition, recall is sometimes also reported over SCF tokens. Token recall gives us the percentage of SCF tokens in the test data that are captured by the SCF dictionary acquired from the training data. To report token recall, we manually built SCFs for 400 verb tokens. If a verb-SCF pair occurs more than once, we only kept one occurrence. This left us with 330 verb-SCF pairs (representing about 130 verb types). out of these 330 verb tokens, the acquired SCF dictionary listed 261 SCFs, giving us a token recall of 79.1%

Verb Tokens	TPs	FPs	False Positive	FNs
ask	9	1		3
begin	2	0		7
believe	2	0		4
cause	2	1	[NP PP(*to*)]	1
expect	3	0		3
find	3	0		5
give	3	0		4
help	6	0		4
like	3	1	[S(that)]	6
move	4	1	[S(that)]	8
produce	2	0		1
provide	3	1	[S(that)]	3
seem	5	0		3
show	5	1	[NP NP]	8
Total	52	5		60
Type precision = 52/(52+5) = 91.2%				
Type recall = 52/(52+60) = 46.7%				

Table 3: Type precision and type recall

One important issue in SCF acquisition, the introduction of high-error-rate components such as taggers and parsers into the acquisition system, is likely to invite skeptics. However, a large scale of SCF acquisition cannot be achieved without using taggers and/or parsers.

Previous SCF acquisition systems that have avoided using taggers and/or parsers have adopted one of the following two alternatives. Sarkar & Zeman (2000) used the Prague Dependency Treebank, which is hand-parsed. However, using hand-tagged and/or hand-parsed text is not a good solution to large-scale SCF acquisition systems as hand-tagging and hand-parsing texts are more laborious than manually collecting SCFs (Manning 1993). Brent (1991, 1992) used lexical cues for identifying verbs as well as SCFs. However, although this procedure was very accurate, this procedure only made use of 3% information of the texts and only extracted a handful of SCF types. As Manning (1993) argued, there are just no highly accurate SCFs cues for many SCFs. The example given in Manning (1993) is the verbs that subcategorize a PP headed by the preposition *in*, such as the ones in (3a, b). However, the majority of occurrences of a PP headed by *in* after a verb are either NP modifiers or non-subcategorized locative phrases, such as those in (3c, d). However, as long as we want to identify verbs that subcategorized for in, we must collect all co-occurrence statistics and use statistical tests to weed out false cues.

(3a) She was assisting the police in that case
(3b) We chipped in to buy her a new TV.
(3c) He built a house in a new surburb.
(3d) We were traveling in a helicopter.

In addition, the noise introduced by parsing errors into SCF acquisition is not as problematic as it appears because not all errors made by taggers and parsers affect the performance of SCF acquisition systems.

First of all, the adoption of the back-off algorithm in hypothesis selection brought some additional benefits to the current system. Sometimes, a SCF cue generated by the Extractor is wrong simply due to an error made by the parser. However, as long as the correct SCF is a subset of the original SCF cue, there exists some chance for the correct SCF to survive the hypothesis selection. For example, the PP *from the Longman Spoken Corpus* in (4) is mistakenly parsed as a verb modifier, leading the Extractor to propose a SCF cue [NP PP(*from*)]. If the false SCF cue fails the hypothesis test, its successor [NP], the correct frame, may still come out as a valid SCF.

(4) … introduce Keith from the Longman Spoken Corpus.

```
    (VP (VB introduce)
         (NP (NNP Keith))
              (PP (IN from)
                   (NP (DT the)
                   (NNP Longman)
                   (VBN Spoken)
                   (NNS Corpus))))))
```

Next, the Extractor does not concern itself with every single level of parsed trees. For example, in (5), the parser wrongly treats adverbs *individually* and *afterwards* as modifying the NP *him*. However, since the adverbs in this example are both adjuncts, the Extractor still manages to propose a correct SCF cue [PP(to)] for *speak* despite of the parsing error.

(5) … speak to him individually afterwards.

```
    (VP (VB speak)
         (PP (TO to)
             (NP (NP (PRP him))
                  (ADVP (RB individually)
                              (RB afterwards))))
```

4 Future work

4.1 Application of the current system to other NLP tasks

The current system should be able to aid several NLP tasks. First, it could probably be incorporated into a parsing system. The current system acquired relative frequency of each SCF that a particular verb takes. Such information can benefit statistical parsers, especially statistical parsers used to parse spoken languages. In addition, we are now interested in the effects of syntactic variables on speech production. This system provides several good candidates for syntactic variables (subjects, objects, complements, and adjuncts).

4.2 Future directions

4.2.1 Some linguistic issues

The current system left some of the linguistic issues unsolved. For example, Manning (1994) pointed out that one of the difficult cases for any English SCF acquisition system is how to represent SCFs for verbs that take a range of prepositional complements (but not all). For example, the verb *put* can take almost any locative or directional prepositional complements. The current system does not have a good way to represent the full range of prepositional complements rather than listing all possible prepositional complements for verbs like *put*. In addition, given that the complement/adjunct distinction is not always clear, sometimes it is difficult to unequivocally determine if a particular constituent is a complement or adjunct. The current system would be of more value if it is able to determine if a particular constituent is more complement-like or adjunct-like.

4.2.2 Hypothesis generation

The current system still showed some room for improvements in hypothesis generation. One way to improve hypothesis generation is to enhance the performance of the parser. However, since parsers are an integral part of any large-scale acquisitions of SCFs and there is an upper limit to how far we could improve parsing accuracy, the improvements we can make on hypothesis generation are limited.

As for the current system, more work could be done to make the Extractor generate SCFs cues with more accuracy. Firstly, the parser does not indicate any complement-gaps in the input sentences. As with the case of passive sentences, the SCF cues proposed by the Extractor will always miss one complement. If the Extractor is made able to correctly identify gaps in the parsed sentences, it will be more accurate in proposing SCF cues. Secondly, the back-off algorithm adopted in this system always removed the last constituent when the original SCF cues are rejected. However, sometimes complements may precede adjuncts. As discussed above, the current system had a good way to deal with adverbs. However, it did nothing to clausal complements. In

English, when a verb takes a clausal complement, the clausal complements tend to extrapose over adjuncts, as in *Mary told John at the party that she was leaving for Chicago tomorrow morning*. The SCF cue proposed by the Extractor for the verb *tell* is [PP(*at*) S(*that*)]. Suppose that this SCF cue is rejected, then the clausal complement S(*that*) will be eliminated. The current Extractor needs to be modified to identify extraposed clausal complements and restore them to a position preceding adjuncts in SCF cues.

4.2.3 Hypothesis selection

It has been shown (Briscoe & Carroll 1997) that hypothesis selection is the weakest link in the acquisition systems of SCFs. As the most popular statistical filtering method, BHT requires estimating error probabilities *p*. However, when estimating the error probabilities *p*, one uses unconditional probabilities *p*(scf) under the assumption that the error probabilities for SCFs are uniform across verbs. For instance, in the current system, the error probabilities for [NP] and [NP S(*that*)] were empirically set to be 0.25 and 0.02 regardless of verbs. Brent (1993) noted that this assumption is false. Most verbs can, for example, take [NP], while very few can take [NP S(*that*)]. As an alternative, we could also use probabilities conditional on verbs *p*(scf|verb). However, this does not work well with SCFs that co-occur infrequently with particular verbs. Korhonen (2002) therefore used probabilities conditional on semantic class *p*(scf|class) based on the observation made by Levin (1993) that semantically similar verbs tend to share the same set of SCFs..

5 Conclusion

Our results show that it is feasible to apply current SCF extraction technology to spoken language. However, it must be admitted that the spoken BNC, which consists mainly of speech in fairly formal settings, is not as challenging as the very informal sociolinguistic interviews of the ViC corpus (Raymond et al, 2002), which have much more disfluency and much more uncertainty about utterance segmentation. Adapting our system to this corpus is our next goal.

6 Acknowledgements

Our thanks go to Eric Fosler-Lussier and David Dowty for their comments. We are also greatly indebted to Mirella Lapata for provding the Charniak- parsed BNC.

References

Brent, M. 1991. Automatic Acquisition of Subcateogrization Frames from Untagged Text. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, Berkeley, CA, 209 – 214.

Brent, M. 1992. Robust Acquisition of Subcateogrization Frames form Unrestricted Text: Unsupervised Learning with Syntactic Knowledge. MS, John Hopkins University, Baltimore, MD.

Brent, M. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics* 19.3: 243 – 262

Briscoe, E. and Carroll, J. 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington, DC. 356 – 363.

Charniak, E. 1997. Statistical parsing with a context-free grammar and word statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence,* AAAI Press/MIT Press, Menlo Park.

Minnen, G., J. Carroll and D. Pearce (2001) ` Applied morphological processing of English', *Natural Language Engineering,* 7(3). 207-223.

Grishman, R., Macleod, C. and Meyers, A. 1994. COMLEX Syntax: Building a Computational Lexicon. In *Proceedings of the International Conference on Computational Linguistics, COLING-94*, Kyoto, Japan. 268-272.

Korhonen A. 2002. Subcategorization Acquisition. Ph.D. dissertation. Computer Laboratory, University of Cambridge.

Lapata, M., and Brew, C. 2004. A Probabilistic Model of Verb Class Disambiguation.

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

Manning, C. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH. 235 – 242.

Raymond, William D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R., and Hilts C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. Proceedings of ICSLP-02. September, 2002. Denver.

Sarkar, A. and Zeman, D. 2000. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 19th International Conference on Computational Linguistics*, Saarbrucken, Germany. 691 – 697.

On Semi-Automatic Detection of Verb Frames and Senses for Ontology Learning

José L. Martínez-Fernández
Computer Science Department
Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés,
Madrid, Spain
jlmferna@inf.uc3m.es

Paloma Martínez
Computer Science Department
Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés,
Madrid, Spain
pmf@inf.uc3m.es

Ana M. García-Serrano
Artificial Intelligence Department
Technical University of Madrid
Campus de Montegancedo s/n, Boadilla del
Monte 28660, Madrid, Spain
agarcia@dia.fi.upm.es

Andreas Nürnberger
IR Group, School of Computer Science,
University of Magdeburg,
Universitätsplatz 2
39106 Magdeburg, Germany
nuernb@iws.cs.uni-magdeburg.de

Abstract

The aim of this paper is to describe an ongoing research approach to the use of verb frames for the semiautomatic extraction of ontologies. The discovering of the frame(s) used for a verb in a corpus can lead to the identification of the elements of an ontology, concepts and relationships among them. This paper presents a possible approach to automatically detect verb senses in free text.

1 Introduction

Since the beginning of the research in the Natural Language Processing (NLP) field there has been an effort devoted to the understanding of written texts. This goal has been so important that a new research field, called Information Extraction, was created. On the other hand, the fast development of the World Wide Web in recent years has lead to a great evolution of the web technology, in particular, for tools and techniques in charge of retrieving accurate information from the web. A part of these advances is the so called Semantic Web, an extended version of the actual web where also conceptual information is represented together with raw content and presentation data. The central element of the Semantic Web is the ontology, "a formal explicit specification of a conceptualization" (Gómez-Pérez et al. 2004) where concepts present on the web are defined and related to each other. There are two main lines on the application of ontologies for content representation: the creation of large well defined and widely used ontologies, and the definition of small application oriented ones. The construction of both types of ontologies requires a great amount of resources, so the development of tools for helping this process has been one of the main targets of the Semantic Web community.

There are several research projects devoted to solving the problem of extracting knowledge or some kind of structured information from texts. Basically, three approaches can be distinguished. The first one, *pattern-based extraction* is centered on conceptual relationships recognized from sequences of words that follow a specific pattern, trying to 'project' linguistic relationships to conceptual or semantic ones; an example of this approach in domain-specific texts is (Kietz et al. 2001). Different sources of information can be considered for this process: *lexical information* is used to detect concepts (in a rough approximation, every noun can become a concept and every adjective can become an attribute for that concept); *syntagmatic information*, by predefined patterns, is used to discover relations between concepts and attributes for them that cannot be detected from lexical information.

In a second approach, the *association rules* framework is used to discover non taxonomic relationships among concepts from a hierarchy of concepts as background knowledge as well as statistics about co-occurrence of words in texts (each pair of closer words with a high co-occurrence can lead to a possible relation between these words; another possibility includes a ngram approach to detect compound words that co-occur in texts), (Maedche and Staab, 2000). Data mining algorithms, devoted to recognizing non-obvious relationships among characteristics in high volume data collections, are then applied.

Finally, the third approach includes *conceptual clustering* where concepts are grouped according to a semantic distance among them, that is, two concepts belong to the same group if their semantic distance is lower than a predefined threshold. One way to calculate the semantic distance among concepts is based on the use of syntactic functions that the terms associated to such concepts play in the text, (Faure and Poibeau, 2000).

The research work introduced in this paper tries to bring together the Information Extraction field, in particular techniques involving verb frames, and the Semantic Web to support the automatic construction of application oriented ontologies.

2 Verb Frames

As stated in (Merlo and Stevenson, 2001), verbs are the basic elements of sentences where relational information among sentence arguments is contained.

A verb frame can be seen as a structure representing, in a formal way, semantic and lexical information of a verb. Syntactic data about verbs (valid arguments, allowed prepositions, etc.) is grouped into subcategorization frames (part of a verb frame), while there is no specific structure to include semantic information such as verb aspect or the corresponding semantic class (Levin, 1993).

The approach described in this paper is based on the assumption that basic semantic information for a verb can be automatically learned from the syntactic level (Merlo and Stevenson, 2001), (Roland and Jurafsky, 1998). If a subcategorization frame of a verb is discovered, it can be used to identify the entities involved in the relation expressed by the verb.

On the other hand, it can also be useful to disambiguate the verb in case of a verb having different senses.

In (Roland and Jurafsky, 1998) some interesting results are described; first, they defend that different verb senses have different subcategorization probabilities and, second, a predefined verb sense has a single subcategorization probability distribution if some factors are controlled, like considering written versus spoken language. So, there is a one to one relation among verb senses and subcategorization frames probabilities. According to these results, if it is possible to detect the different subcategorization frames of a verb, the used verb sense can be identified and a relation among different entities present in the text can be automatically identified.

In the previous description, there is a missing step, how can we make the mapping among subcategorization frames and verb senses. For this purpose, different hand coded semantic resources can be used such as FrameNet (Fillmore et al., 2002) or VerbNet (Kipper et al., 2000).

There have been a lot of previous research works where statistical and linguistic information is combined to obtain and learn verb frames from text like those contained in (Klein and Manning, 2004), (Zeman and Sarkar, 2000), (Maragoudakis et al., 2000), (Korhonen 1998), (Chen and Chen, 1994), (Manning, 1993).

2.1 Syntactic guided detection of verb frames

The final goal pursued by the work described in this paper is to automatically extract relations among entities by a frequency of appearance analysis of the syntactic structures in which a verb appears, given an English texts collection. The process followed is described bellow.

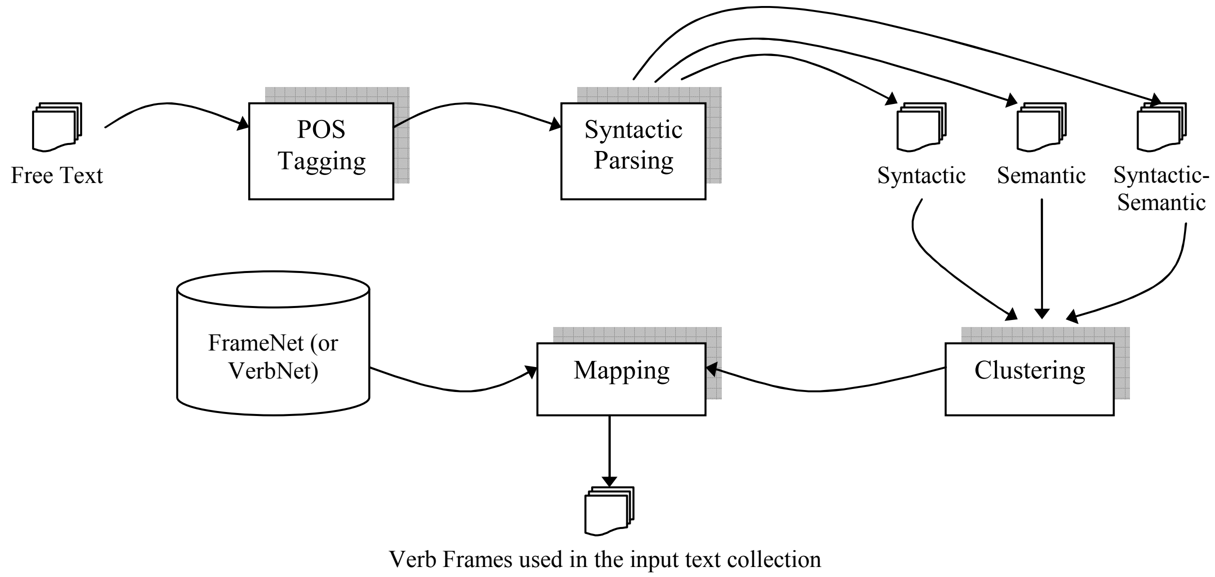


Figure 1. Defined process to extract verb senses from free text

First, the text is passed through a POS tagger, the Brill tagger (Brill, 1992) in the first developed experiments.

Second, a parser is used (Charniak, 1997) to recognize syntactic structures, like noun phrases and verb groups. The result of this syntactic analysis is used to identify the different verbs appearing in the texts collection.

Third, data structures surrounding the different occurrences of a verb are extracted to act as the input of the next step.

Three levels of linguistic information can be distinguished to build the mentioned data structures:

- *Semantic level*, where only words are considered. For example, for the verb fall, possible occurrences can be: "copper stocks fall in January" or "trade deficit falls in January".
- *Syntactic level*, where only syntactic structures around the verb are used to represent the verb occurrence. For example, for the verb fall, possible syntactic occurrences could be: "[<NP><VP>]", a noun phrase followed by a verb phrase containing the specified verb.
- *Syntactic-semantic level*, where each syntactic element is accompanied by its core word (i.e. the first name in the case of a noun phrase). For example, for the verb fall, a possible occurrence could be: "[<VBG> [<IN-by> <NP;pct>]]"

These three different linguistic levels are introduced to answer the following question: how much information is needed to recognize the frame used in an occurrence of a verb? or, in a more realistic expression, can any of this levels lead to the identification of the frame used in an occurrence of a verb?

Fourth, the list obtained in the previous step is used to apply (preferably unsupervised) clustering techniques. This process produces several clusters, each of them represented by the involved syntactic elements.

Fifth, the obtained clusters are mapped against the different frames supplied for the verb by external resources like FrameNet or VerbNet. In the case of FrameNet, this mapping could be carried out by representing the identified cluster and the different lexical frames in FrameNet using vectors; the distance among these vectors could point the applied verb frame.

At the end of this process, the different verb senses used in the text can be recognized and elements linked by the verb can be identified along with their roles in the verb structure.

2.2 Preliminary experiments

Some preliminary experiments are being carried out using as input part of the Reuters-21578 test collection (Lewis, 1997). This test collection is analyzed using the Brill tagger and Cherniak's parser. Three different outputs, according to the previously described linguistic levels, are generated. For this purpose, several software tools have been developed using the Java programming language, which operate over an XML representation of the tagged and parsed documents. The frequency of appearance of each frame occurrence is collected and the size of context of the verb is a parameter of the software tools. These outputs are clustered using WEKA library (from www.cs.waikato.ac.nz/ml/weka). In a first attempt, simple k-means and EM clustering algorithms are taken into account, although other possibilities are being analysed. Unfortunately, at the time of writing we are unable to provide conclusive results. However, so far, the first results are very promising.

Semantic Frames	
Frame	**Frequency**
FALLING 0.6 pct	2
FORECASTS PROFITS FALL IN 1987 / 88	2
to FALL to 328	2
the FALL was largely	2
FALLS NEW YORK	2
PROVEN RESERVES FALL FORT WORTH	2
3.4 pct FALL for passenger	2
car sales FELL one pct	2

Table 1. Examples of recognized semantic frame occurrences for the verb "fall" in the Reuters collection.

Syntactic Frames	
Frame	**Frequency**
[[<NP><VP><PUNCT>]]	35
[[<NP><VP>]<PUNCT>[<NP><VP>]]	27
[[<NP><VP>]]	26
[<IN-after>[<VP>]]	14
[[<TO><VP>]]	8
[[<NP><VP>]<PUNCT>[<NP><VP>]< PUNCT>]	6
[<IN-that>[<NP><VP>]]	5
[[<NP><VP>]<PUNCT>[<PRP>][<VB D>]<PUNCT>]	3

Table 2. Examples of recognized Syntactic frame occurrence for the verb "fall" in the Reuters collection.

Syntactic-Semantic Frames	
Frame	**Frequency**
[<DT><NN;fall>]	16
[<JJ><NN;Fall>]	7
[<VBD>[<QP><NNS;dlrs>][<TO><NP; dlrs>]]	5
[<DT><JJ><NN;fall>]	5
[<VBD>[<NP;dlrs><PUNCT><CC-or><NP;pct><PUNCT>][<TO><NP;dlrs>]]	4
[<VBD>[<CD><NN;pct>][<TO><NP;units>][<IN-from><NP;units><NP;year>]]	4
[<DT>[<CD><NN;pct>]<NN;fall>]	3
[<VBD>[<CD><NN;pct>]]	3

Table 3. Examples of recognized syntactic-semantic frame occurrence for the verb "fall" in the reuters collection.

In the following tables some examples of each kind of linguistic frame occurrence are shown. The verb "fall" has been selected to extract these examples of frames occurrences.

The final result of the system is a vector representation of the recognized clusters. This vector has frame elements as components, which are accompanied by a weight.

3 Ontologies

Although the ontology concept is used in quite diverse and sometimes even misleading ways, for the purpose of this article an ontology is considered as a description of a shared conceptualization. That is, a vocabulary of terms and relationships among them that have to be defined in a formal and machine readable format. Since ontologies are used for different purposes (natural language processing, Semantic Web, etc.), in different disciplines (artificial intelligence, databases, etc.) and in specific application domains, ontologies may adopt a variety of forms but the terms and relationships require some formal and explicit specification of their semantics in order to restrict the interpretation on a specific domain.

The On-To-Knowledge methodology, (Staab et al., 2001) proposes four main processes for ontology development: Process 1: *feasibility study*, process 2: *kickoff*, process 3: *refinement*, process 4: *evaluation* and process 5: *maintenance*.

Process 2 is devoted to describing domain and goal of the ontology, the design guidelines, available resources, potential users and use cases as well as applications supported by the ontology; Process 3 has as goal to produce an application oriented ontology according to the specification obtained in Process 2. Process 3 is divided into two activities: Activity 1: *knowledge elicitation process with domain experts* and Activity 2: *formalization*. In Activity 1 the draft of the ontology obtained in Process 2 is refined by identifying and modeling axioms with experts in the domain. During the elicitation, the concepts are gathered on one side and the terms to label the concepts on the other. Then, terms and concepts are mapped. Precisely in this activity, a way of enriching the ontology is to apply a *learning* process to increase the number of concepts and associations among them.

4 Application of automatically detected subcategorization frames for ontology learning

As described in the previous section, the basic elements of an ontology are the entities representing the concepts of the problem domain and the existing relations among these concepts. (Maadche and Staab, 2000) distinguish among different kinds of relations that can be extracted from texts, i.e. taxonomic and not taxonomic relations. Usually, taxonomical relations can be extracted using clustering techniques but non taxonomic relations are more difficult to recognize. For this purpose, verb frames can be used (Faure and Nédellec, 1999). The verb involved in a frame gives the type and name of the relation, while the arguments of the frame would serve as indicators of the concepts to be linked by the verb relation. Following these directions, if the verb frame can be automatically extracted, also concepts and non taxonomic relations among them can be automatically identified. Of course, there is certain error rate introduced in the process, mainly related with tagging and parsing errors but, of course, also due to the gap between syntactic and semantic frames.

Taking into account this process, the goal of this research work is to build a system that can be integrated in an ontology learning environment. This system would serve as a tool for semi-automatically extract non taxonomic relations among concepts from free text.

5 Conclusion

The approach described in this paper seems to be a feasible way to perform automatic verb sense extraction from free text, although, to the time of writing, no conclusive experimental results can be reported. Some difficulties related to the detection of verb occurrences have been recognized during the development of the text analysis tools, which will be fixed in future versions. Besides, the mapping process to match the recognized verb frame with external semantic resources (like FrameNet or VerbNet) is difficult to define and, perhaps, it must be defined as a human-aided process.

The effect of tagging and parsing errors has not been forgotten. The use of statistical based tools to obtain the tagged and parsed versions of texts can introduce an important percentage of errors. In the proposed approach, the final goal is not to obtain the correct frame for every verb in the text collection but it is enough to extract the right probabilistic distribution of verb frames into the collection. In this way, we believe that errors produced by the automatic tagging and parsing of the input text could be ignored.

As mentioned in the introduction, this is just an ongoing research work and, so, there is a lot of work to do. In a first step, the best suited clustering algorithm must be identified and, as a second step, an automatic mapping among external semantic resources and artificially built clusters must be developed. Once the process is developed, it must be integrated in an ontology learning system, which will be used as a framework to evaluate the verb sense recognition module.

6 Acknowledgements

Part of this work has been carried out while José L. Martínez-Fernández, one of the authors of this paper, was at the University of Magdeburg as a Visiting Researcher of the Information Retrieval Group.

This work has been partially supported by the projects OmniPaper (European Union, 5th Framework Programme for Research and Technological Development, IST-2001-32174), MIRACLE (Regional Government of Madrid, Regional Plan for Research, 07T/0055/2003) and GPS Project – Software Process Management Platform: modeling, reuse and measurement (National Research Plan, TIN2004-07083).

References

Brill E. 1992. *A simple rule-based part of speech tagger*. Proceedings of the Third Conference on Applied Natural Language Processing, ANPL, ACL, Trento, Italy, pp. 152-155.

Charniak, E. 1997. *Statistical parsing with a context-free grammar and word statistics*, Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI Press/MIT Press.

Chen K. and Chen H. *1994. Acquisition of Subcategorization Frames from Large Scale Texts.* In Proceedings of the Second Conference for Natural Language Processing (KONVENS-94), Vienna, Austria, September 28-30, 1994, pp. 407-410.

Faure D. and Nédellec C. 1999. *Knowledge acquisition of predicate argument structures from technical texts using Machine Learning: the system ASIUM*. In Proceedings 11th European Workshop EKAW'99, pp. 329-334.

Faure and Poibeau, 2000, First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEXT. In: Staab S, Maedche A, Nedellec C. Wiemer-Hastings P. eds. Ontology Learning ECAI-2000 Workshop, pp 7-12.

Fillmore, C. J., Baker, C. F. and Sato, H. 2002. *The FrameNet Database and Software Tools*. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), pp.1157-1160.

Gómez-Perez A., Fernández M. and Corcho O. 2004, *Ontological Engineering*, Springer.

Kietz J., Maedche A. and Volz E. 2000. A method for semi-automatic ontology acquisiton from a corporare Intranet. In: Aussenac-Gille N, Biébow B, Szulman S. eds. EKAW'00 Workshop on Ontologies and texts.

Kipper K., Dang H. T. and Palmer, M. 2000 *Class-Based Construction of a Verb Lexicon*. AAAI-2000.

Klein D. and Manning C. D. 2004. *Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency.* In Proceedings *Proceedings of the 42nd Annual Meeting of the ACL*

Korhonen A. *1998. Automatic Extraction of Subcategorization Frames from Corpora - Improving Filtering with Diathesis Alternations.* In Proceedings of the ESSLLI 98 Workshop on Automated Acquisition of Syntax and Parsing. pp. 49-56.

Levin, B. 1993. English Verb Classes and Alternations. University of Chicago Press, Chicago, IL.

Lewis, D. 1997. Reuters-21578 text categorization test collection.

Maedche A. and Staab S. 2000, Mining Ontologies from texts. In: Dieng, R, Corby, O eds. 12th International conference in Knowledge Engineering and Knowledge Management (EKAW'00), LNAI 1937, pp 189-202.

Manning C. D. *1993, Automatic acquisition of a large subcategorization dictionary from corpora.* In Proceedings of the 31st conference on

Association for Computational Linguistics, pp. 235-242.

Maragoudakis M., Kermanidis K. L. and Kokkinakis G. 2000. *Learning Subcategorization Frames from corpora: A case Study for modern Greek.* In Proceedings of COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries, pp. 19-22, Kato Achaia, Greece, 22-23.

Merlo P. and Stevenson S. 2001. *Automatic Verb Classification Based on Statistical Distributions of Argument Structure.* In Computational Linguistics, Vol. 27, N. 3.

Roland D. and Jurafsky D. *1998. How Verb Subcategorization Frequencies Are Affected By Corpus Choice.* In Proceedings of the 17th international conference on Computational linguistics - Volume 2, pp.1122-1128.

Staab S., Schnurr HP, Studer R. and Sure Y. 2001. Knowledge Processes and Ontologies. IEEE Intelligent Systems 16, 1, pp.26-34.

Zeman D. and Sarkar A. *2000. Learning Verb Subcategorization from Corpora: Counting Frame Subsets.* In 2nd International Conference on Language Resources and Evaluation (LREC2000).

Unaccusative/Unergative Distinction in L2 English by Spanish and Japanese Native Speakers

Keiko MATSUNAGA
Department of Language & Linguistics, University of Essex
Wivenhoe Park, Colchester
Essex CO4 3SQ U.K.
kmatsup@essex.ac.uk

Abstract

It has been claimed that second language (L2) learners are sensitive to the distinction between unaccusative and unergative verbs. Central issues to be addressed in the present study are (i) whether this lexical distinction is observed in L2 grammars regardless of a native language (L1) and (ii) whether there is any effect of L1 transfer in L2 acquisition of unaccusative/unergative verbs. In order to answer these questions, this study investigates the L2 acquisition of transitivity alternations with three intransitive verb classes (alternating unaccusative verbs, non-alternating unaccusative verbs, and non-alternating unergative verbs) by Spanish- and Japanese-speaking adult L2 learners of English. The results of an acceptability judgement task observe both L1 transfer as well as the universal distinction of unaccusative/unergative verbs at the level of lexical argument structure proposed by Hale and Keyser (2002).

1 Introduction: Unaccusative vs. Unergative

English intransitive verbs can be categorized into the following three sub-classes:

(1) Unaccusative verb
a. The window broke.
b. Bill broke the window.
c. The window was broken (by Bill).
d. Bill made the window break.

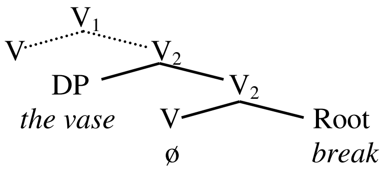
(2) Non-alternating Unaccusative verb
a. A rabbit appeared.
b. *The magician appeared a rabbit.
c. *A rabbit was appeared (by the magician).
d. The magician made a rabbit appear.

(3) Non-alternating Unergative verb
a. The students laughed.
b. *The teacher laughed the students.
c. *The students were laughed (by the teacher).
d. The teacher made the students laugh.

Alternating unaccusative verbs as in (1) allow both intransitive and transitive variants. Passivisation and causativisation with the analytic causative *make* are also possible, but the acceptability will depend on the discourse context. The passive form is typically preferred over the simple intransitive verb form when there is an explicit or implied agent. The simple intransitive is typically preferred when there is no identifiable agent. The analytic causative form *make* is typically preferred over the simple transitive when the emphasis is on the agent's indirect involvement in the event. Conversely, the other two verb classes do not have transitive alternants, but require *make* to form a causative as in (2) and (3). Thus, these verb classes can be neither passivised nor simply transitivised without a causative verb.

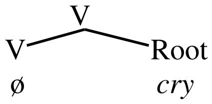
The theoretical framework adopted in this study is Hale and Keyser (2002). They assume that each verb consists of two structural elements; a root (R) and a verbal host (V) as shown in (4). In the case of the unaccusative verb *break*, it has a dyadic structure involving an internal argument in the specifier of the verb. Adding an additional V layer (V_1 in the tree) allows this verb to alternate.

(4) Lexical configuration of unaccusative verbs



Conversely, in the case of unergative verbs, the lexical configuration is a simple monadic structure which consists of Head and Complement only as illustrated in (5).

(5) Lexical configuration of unergative verbs



The lexical configuration of the unergative verb has no specifier which is a place for an internal subject. Lack of an internal argument disallows the structure (5) to merge with an additional V to form a transitive variant, and thereby unergative verbs always require an external subject.

Although it has been believed that the distinction between unaccusative and unergative verbs are universal at the level of lexical argument structure, cross-linguistic differences may arise in terms of the morphosyntactic realisation, as will be discussed in section 2.

2 Cross-linguistic Differences in Morphosyntactic Realisation

As compared with the English examples presented in (1-3), this section discusses cross-linguistic differences in the morphosyntactic realisation of three verb classes between Spanish and Japanese.

2.1 Spanish

Spanish obligatorily marks the intransitive variant of the alternating unaccusative verbs with the reflexive clitic *se* as in (6a), whilst its transitive form has no morphological marking like English as in (6b). Most unaccusative verbs in Spanish optionally take *se* as in (7a). In the case of unergative verbs, some are marked with *se* as in (8a), but others do not as in (8b). The non-alternating verb classes in Spanish lack a transitive alternant and so require the analytic causative *hacer* as in (7b) and (8c).

(6) Alternating Unaccusative verbs in Spanish
a. *El vaso* ***se*** ***rompió***.
the vase reflexive clitic break-past
"The vase broke."
b. *Bill* ***rompió*** *el vaso*.
Bill break-past the vase
"Bill broke the vase"

(7) Non-alternating Unaccusative verbs in Spanish
a. *Un conejo* ***(se) desapareció***.
a rabbit disappear-past
"A rabbit disappeared"
b. *El mago* ***hizo*** ***desaparecer*** *el conejo*.
the magician made disappear the rabbit
"The magician made the rabbit disappear"

(8) Non-alternating Unergative verbs in Spanish
a. *Susie* ***se*** ***sonrió***.
Susie reflexive clitic smile-past
"Susie smiled"
b. *Mary* ***bailó***.
Mary dance-past
"Mary danced"
c. *El fotógrafo* ***hizo*** ***sonreir*** *a Susie*.
the photographer made smile Susie
"The photographer made Susie smile"

2.2 Japanese

The intransitive and transitive variants of the alternating unaccusative verbs are morphologically related in Japanese as in (9). Interestingly, the verbs which are categorized into the non-alternating verbs in English, such as *disappear*, can alternate through morphological derivation in Japanese as in (10) (e.g. *ki-e-ru* [intr.] / *ke-su* [tr.]). Like English and Spanish, the Japanese unergative verbs cannot have a transitive variant and the causative morphology is required as in (11b).

(9) Alternating Unaccusative verbs in Japanese
a. *Kabin-ga* ***kow-are****-ta*.
the vase-NOM break-Intr-past
b. *Bill-ga* *kabin-o* ***kow-asi****-ta*.
Bill-NOM the vase-ACC break-Tr-past

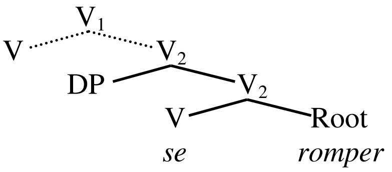
(10) Non-alternating Unaccusatives in Japanese
a. *Usagi-ga* ***ki-e****-ta*.
rabbit-NOM disappear-Intr-past
b. *Tejinashi-ga* *usagi-o* ***ke-si****-ta*.
magician-NOM rabbit-ACC disappear-Tr-past

(11) Non-alternating Unergative verbs in Japanese
a. *Susie-ga* ***hohoen****-da*.
Susie-NOM smile-past
b. *Kameraman-ga* *Susie-o* ***hohoem-ase****-ta*.
photographer-NOM Susie-ACC smile-Caus-past

2.3 Morphology in lexical argument structure

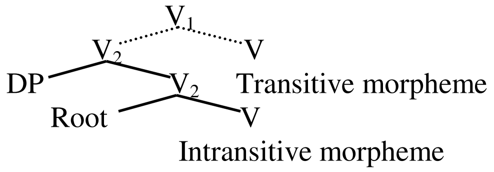
Within the framework of Hale and Keyser, these cross-linguistic differences result from the presence or absence of morphological components in the verbal heads. In the case of Spanish, for instance, the anticausative morphology *se* occupies the lower verbal head as illustrated in (12).

(12) Spanish unaccusative verbs



In Japanese, on the other hand, the overt in/transitive morphemes associated to the alternation occupy the upper and lower verbal heads respectively.

(13) Japanese unaccusative verbs



Thus, the relevant morphological differences across languages correlate directly with a structural difference at the level of lexical argument structure. The morphological reflexes are manifestations of structural differences, not the cause of those differences.

3 Unaccusative vs. Unergative in L2

Much attention has been paid to these three classes of intransitive verbs in acquisition studies. Hirakawa (1995) examined the distinction between unaccusative and unergative verbs in L2 by testing the intermediate Japanese-speaking learners of English. The result of her judgement task showed that learners tended to accept incorrect passive form more with non-alternating unaccusative verbs (e.g. **John was fallen down*) than the unergative verbs (e.g. **Bill was cried*).

Cabrera and Zubizarreta (2003) investigated the knowledge of the distinction between two non-alternating verb classes by testing ungrammatical simple transitive constructions. They found that the English-speaking learners of Spanish at the low proficiency level incorrectly accepted simple transitive significantly more with unaccusatives (e.g. **El padre llegó a la niña tarde* "*The father arrived the girl late") than unergatives (e.g. **El payaso rió al niño* "*The clown laughed the boy").

The previous studies provide evidence for the unaccusative/unergative distinction in L2 grammars. Assuming that the unaccusative verbs have a dyadic structure (4), compared with the unergative verbs which have a monadic structure (5), it incorrectly leads learners to overgeneralise passive and simple transitive form with the non-alternating class of the unaccusative verbs.

Hypothesis:
Given that the unergative and unaccusative verbs have distinct lexical configurations universally, learners make overgeneralisation errors only with the unaccusative verbs, regardless of their L1s.

4 Experimental Study

28 Spanish native speakers and 27 Japanese native speakers participated in this study with 14 English native speakers serving as control. All the subjects were recruited at a university in England and paid for their participation. The Placement Test (University of Cambridge Local Examinations Syndicate, 2001) was used as an independent measurement for evaluating their English proficiency. Based on a mean score of the QPT (43.44, Maximum score=60, SD=6.65) they were divided into two levels (Lower vs. Upper proficiency level) in each group.

Language group	English proficiency levels	Mean QPT scores (SD)
English	Control (n=14)	-
Spanish	Lower (n=13)	37.31 (3.52)
	Upper (n=15)	49.00 (4.46)
Japanese	Lower (n=12)	37.17 (2.82)
	Upper (n=15)	48.20 (2.43)

Table 1: Subjects

An acceptability judgement task was used as a main task in this study. Twelve verbs in three different verb classes were involved.

Alternating unaccusatives	Non-alternating unaccusatives	Unergatives
break melt close bend	happen disappear arrive occur	laugh dance smile walk

Table 2: Verbs tested in the task

In this task each question contained one short introductory sentence and two continuations. One of the continuations had a morphologically overt verbal form (e.g. *The vase was broken*; *Bill made the glass break*) and the other had a morphologically zero verbal form (e.g. *The vase broke*; *Bill broke the glass*). Different contexts were provided to elicit each form; the "non-passive context" in (14) was intended to elicit the simple intransitive verbal form over the *be*-passive forms, while the "passive context" in (15) was intended to elicit the *be*-passive form.

(14) *That vase had been cracked since I dropped it last Christmas. Yesterday finally…*
a. The vase broke.
b. The vase was broken.

(15) *While washing dishes after the dinner, Tom dropped one of the plates.*
a. The plate broke.
b. The plate was broken.

For testing transitive constructions, the "direct causative context" in (16) was intended to elicit the simple transitive verbal form, while the "indirect causative context" in (17) which involves an instrumental agent, was intended to elicit the analytic *make* causative form.

(16) *Bill got this glass at a low price. When he squeezed it too hard, however, …*
__a. Bill broke the glass.__
b. Bill made the glass break.

(17) *Bill found a new glass left in the kitchen. When he poured boiling water into it, …*
a. The heat broke the glass.
__b. The heat made the glass break.__

The questions were presented on a computer screen and subjects were asked to choose an answer from three options, "impossible", "possible" and "natural". If the subject could not decide the answer, another option, "not sure", was chosen, and it was excluded in the analysis. Before testing, the test instructions and some examples were presented to make sure that they should judge each sentence in terms of both the meaning of the sentence and the correctness of the grammaticality.

5 Results

5.1 Intransitive constructions

This section presents results of the intransitive constructions. In the analysis, each answer is calculated as follows; Impossible=0, Possible=1, and Natural=2. Since the distribution of the data was not normal, nonparametric statistics have been used. Significant differences found are indicated in tables as follows: **= p<.01, *= p<.05.

Table 3 shows the mean acceptability rates of the simple intransitive verbal form (e.g. *The vase broke*) and passive form (e.g. *The vase was broken*) with alternating unaccusative verbs.

Lang. group	Non-Passive contexts		Passive contexts	
	simple intransitive	*be*-passive	simple intransitive	***be*-passive**
E	**1.91****	.71	1.02	**1.39***
SL	**1.10**	1.38	.49	**1.59****
SU	**1.78****	1.02	.91	**1.70****
JL	**.85**	1.67**	.44	**1.87****
JU	**1.29**	1.30	.80	**1.58****

Table 3: Alternating unaccusative verbs

The English control (E) group correctly distinguished between the non-passive contexts and the passive contexts; significant differences in their acceptability rates between the simple intransitive verbal forms and the morphologically overt verbal forms are found in both contexts. The Spanish upper proficiency group (SU) behaved similarly to the control group. However, the learners at the lower proficiency levels in both Spanish (SL) and Japanese (JL) groups displayed a tendency to accept the passive forms regardless of the contexts. The JL group in particular showed a significant preference for the passive form even in the non-passive contexts. The Japanese learners at the upper proficiency levels (JU) did not show any preference in the non-passive contexts.

Table 4 shows the results of the non-alternating unaccusative verbs.

Lang. group	Non-Passive contexts		Passive contexts	
	simple intransitive	**be*-passive	**simple intransitive**	**be*-passive
E	**2.00****	.00	**1.93****	.00
SL	**1.92****	.33	**1.87****	.42
SU	**1.98****	.13	**1.97****	.17
JL	**1.69****	.81	**1.60***	1.10
JU	**2.00****	.22	**1.90****	.45

Table 4: Non-alternating unaccusative verbs

In this verb class, the simple intransitive verbal form (e.g. *The car accident happened*) is only available, whilst the passive form (e.g. **The car accident was happened*) is ungrammatical. However, errors with the overuse of passives can be found among L2 learners, especially in the JL group; the results imply that they judged the ungrammatical sentences as "possible" (.81 in the non-passive contexts and 1.10 in the passive contexts).

Table 5 represents the results of non-alternating unergative verbs. Since this verb class does not have the transitive alternant, the passive form (e.g. **The child was cried*) is unavailable and the simple intransitive verbal form is only possible (e.g. *The child cried*).

Lang. group	Non-Passive contexts		Passive contexts	
	simple intransitive	**be*-passive	**simple intransitive**	**be*-passive
E	**1.98****	.02	**1.98****	.00
SL	**1.88****	.12	**1.88****	.13
SU	**1.98****	.00	**1.98****	.07
JL	**1.90****	.56	**1.77****	.73
JU	**1.97****	.20	**1.93****	.27

Table 5: Non-alternating unergative verbs

Although all groups displayed a significant preference for the simple intransitive verbal form and errors with the overuse of passives were rarely found with this verb class, the JL group behaved differently from all the other groups; the JL group was the only group that showed an apparent overuse of the ungrammatical passive forms (.56 in the non-passive contexts and .73 in the passive contexts).

5.2 Transitive constructions

Let us now turn to the result of transitive constructions. The results of a comparison between the simple transitive verbal form (e.g. *Bill broke the glass*) and the *make* causative form (e.g. *Bill made the glass break*) with the alternating unaccusative verbs are presented in Table 6.

Lang. group	Direct causative contexts		Indirect causative contexts	
	simple transitive	*make* causative	simple transitive	***make* causative**
E	**1.98****	.71	1.64	**1.66**
SL	**1.96****	.58	1.67**	**1.15**
SU	**1.93****	.83	1.65*	**1.31**
JL	**1.83****	.89	1.53	**1.46**
JU	**1.92****	.66	1.54*	**1.18**

Table 6: Alternating unaccusative verbs

A significant preference for the simple transitive verbal form in the direct causative contexts was observed in the control group as well as all the experimental groups. Although the difference in the acceptability rates between the simple transitive verbal form and *make* causative is not significant in the indirect causative context, compared to the results of the direct causative context, the acceptability rate of the *make* causative significantly increases in all the groups.

Table 7 shows the results of the non-alternating unaccusative verbs. In this verb class, the simple transitive verbal form (e.g. *_John happened the accident_) is ungrammatical. Instead, the analytic causative construction with *make* (e.g. *John made the accident happen*) is required.

Lang. group	Direct causative contexts		Indirect causative contexts	
	*simple transitive	***make* causative**	*simple transitive	***make* causative**
E	.00	**1.09****	.00	**1.36****
SL	.60	**1.51****	.53	**1.74****
SU	.36	**1.63****	.36	**1.77****
JL	.94	**1.35***	.90	**1.38***
JU	.45	**1.22****	.37	**1.54****

Table 7: Non-alternating unaccusative verbs

The English native speakers completely rejected the ungrammatical sentences. Conversely, the learners tended to overgeneralise the incorrect simple transitive verbal form with this verb class. The learners in the JL group especially judged it as "possible" (.94 in the direct causative contexts and .90 in the indirect causative contexts).

Contrary to the results of the non-alternating unaccusative verbs, the learners correctly rejected the ungrammatical sentences with non-alternating unergative verbs as shown in Table 8. In this verb class, the simple transitive verbal form (e.g. *_The clown laughed the children_) is unavailable and the *make* causative is always required (e.g. *The clown made the children laugh*).

Lang. group	Direct causative contexts		Indirect causative contexts	
	*simple transitive	***make* causative**	*simple transitive	***make* causative**
E	.05	**1.87****	.00	**1.80****
SL	.10	**1.94****	.04	**1.76****
SU	.14	**1.97****	.03	**1.90****
JL	.31	**1.83****	.25	**1.69****
JU	.28	**1.93****	.20	**1.77****

Table 8: Non-alternating unergative verbs

All the experimental groups showed a reluctance to accept the ungrammatical sentences with the unergative verbs. Additionally, statistical analysis shows significant differences in the acceptability of the incorrect simple transitive verbal forms between two non-alternating verb classes (non-alternating unaccusative verbs in Table 7 vs. unergative verbs in Table 8) in the SU group (p≤.047), the SL group (p≤.002), and the JL group (p≤.001).

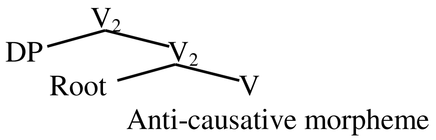
6 Discussion

Let us summarise the results in terms of the hypothesis. The results of the intransitive constructions showed that the learners at the upper proficiency levels overused passive forms with the alternating unaccusative verbs, but hardly ever with the non-alternating classes. The lower proficiency groups showed a tendency to accept the ungrammatical passive forms more with the non-alternating unaccusative verbs than the unergative verbs. As for the results of transitive constructions, although the learners tended to accept ungrammatical simple transitive verbal forms with the non-alternating unaccusative verbs, these errors were rarely found with the unergative verbs. Such results are consistent with the previous studies which showed the unaccusative/unergative distinction in L2 acquisition. Simultaneously however, a cross-linguistic difference between Japanese and Spanish groups has been observed in the present study; the Japanese speakers made more errors than the Spanish speakers, accepting both ungrammatical passive forms and simple transitive verbal forms with non-alternating verb classes.

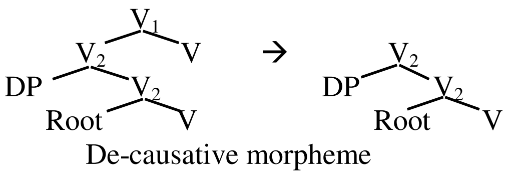
In order to explain such a cross-linguistic difference, let us now look at the derivation of intransitive verbs in two languages more closely. According to Kageyama (1996), there are two processes of intransitivisation in Japanese; one is "anti-causativisation" in which a causer is identified with a causee, and another is "de-causativisation" in which a causer becomes unspecified and an internal argument surfaces as a sentential subject. Furthermore, Japanese has morphology which attaches to an intransitive verb and derives a transitive form; the non-alternating unaccusative verbs in Japanese take this morphology so that the intransitive verb *ki-e-ru* "disappear" does alternate through this process of "causativisation". Notice that English and Spanish allow the anti-causativisation only. Thus, compared to these two languages, the transitivity alternation is available with a wide range of verbs in Japanese due to a variety of morphemes.

We assume that the anti-causative morphology blocks transitivisation by checking the Accusative Case as in (18), while the de-causative morphology attaches to the transitive alternant, and then removes the upper verbal head, leaving the lower verbal head, namely its intransitive variant, as shown in (19).

(18) Anti-causativisation



(19) De-causativisation



Although these two lexical argument structures arrive at an identical surface structure for the intransitive variants (except for the distinct morphemes which occupy the lower verbal head), the underlying derivational processes are different. In other words, they have different conflation patterns. Given that the process of de-causativisation derives from a transitive variant, this structure always involves the initial presence of an agent. Thus, Japanese speakers may assume that English has the processes of de-causativisation and causativisation (even though it only has anti-causativisation) at least at the lower proficiency level; this would explain the reason why the Japanese speakers tended to accept more incorrect passive and simple transitive verbal forms than the Spanish speakers. If this analysis is on the right track, L1 transfer of the morphology is not enough to explain the Japanese speakers' behaviour and we would consider that the whole lexical argument structure which involves the conflation pattern and the argument changing morphology is a target of L1 transfer.

7 Conclusion

To conclude, the overall results suggest that unaccusative and unergative verbs are represented differently in L2 grammars; unaccusatives have a dyadic structure, whilst unergatives have a monadic structure. However, the present study has observed a significant effect of L1 transfer as well. Even though the lexical configuration is universal across languages, different derivational patterns attributed to a variety of morphemes and different conflation patterns led the Japanese learners of English to overuse both incorrect passive and simple transitive verbal forms with non-alternating verb classes more than the Spanish learners of English in this study.

References

M. Cabrera. and M. L. Zubizarreta. 2003. *On the acquisition of Spanish causative structures by L1 speakers of English.* In "Proceedings of the 2002 Generative Approaches to Second Language Acquisition (GASLA-6): L2 Links", J. M. Liceras, H. Zobl and H. Goodluck, ed., pages 23-33, Cascadilla Press, Somerville, Massachusetts.

K. Hale. and S. J. Keyser. 2002. *Prolegomenon to a Theory of Argument Structure*. The MIT Press, Cambridge, Massachusetts.

M. Hirakawa. 1995. *L2 acquisition of English unaccusative constructions.* In "Proceedings of the Boston University Conference on Language Development (BUCLD) 19", D. MacLaughlin and S. McEwen, eds., pages 291-302, Cascadilla Press, Somerville, Massachusetts.

T. Kageyama. 1996. *Doshi Imiron* ("*Verb Semantics*"). Kurosio publishers, Tokyo.

University of Cambridge Local Examinations Syndicate. 2001. *Quick Placement Test.* Oxford University Press, Oxford.

Automatic Learning of Syntactic Verb Classes

Laia Mayol and **Gemma Boleda** and **Toni Badia**
GLiCom
Dept. of Translation and Philology
Pompeu Fabra University
{laia.mayol,gemma.boleda,toni.badia}@upf.edu

Abstract

This paper describes an experiment devised to group Catalan verbs according to their syntactic behavior. Our goal is to acquire a small number of basic classes with a high level of accuracy, from relatively knowledge-poor resources. This information, expensive and slow to compile by hand, is useful for any NLP task requiring specific lexical information.

The experiment aims at automatically classifying verbs into transitive, intransitive and verbs alternating with a *se*-construction. We use a clustering methodology applied to data extracted from a tagged corpus. Our system achieves an average 0.87 F-score, for a task with a 0.65 baseline. The cluster analysis also provides insight into the relevant features and the notion of prototypicality within a class.

1 Introduction

This paper presents a method to automatically classify Catalan verbs into syntactic classes by means of clustering, an unsupervised machine learning technique. Obtaining lexical information about the linguistic behavior of every word is critical for many NLP tasks, specially in the case of verbs, as they have a great influence in the syntactic pattern and the informational content of the sentence.

However, manually compiling this information is an expensive and slow task, which is never complete and often leads to inconsistent resources (Ide and Véronis, 1998). In the last decade, much research has focused on lexical acquisition, that is, on inferring lexical properties of words from their behavior in corpora and other resources, using machine learning techniques.

The first works on automatic acquisition of subcategorisation information were not directed at classifying verbs but at compiling every possible subcategorisation frame for each verb. Brent (1993) used raw corpora to obtain six different frame types; Manning (1993) described a system which could recognize up to 19 frames; Briscoe and Carroll (1997) followed this same line but dealt with 160 frames. Several works in recent years are closer to the goals or methodology of the experiments presented in this paper. Merlo and Stevenson (2001) applied supervised techniques to acquire three different classes of optionally transitive verbs: unergative, unaccusative and object-drop. They achieved 69.8% accuracy. The technique we use here, clustering, has also been used to classify verbs into semantic classes (Schulte im Walde, 2000; Stevenson and Joanis, 2003).

The approach in this and other related work is to use (mainly) syntactic features to induce semantic classes, thus exploiting the syntax-semantics interface. Our task is arguably simpler, because it uses syntactic cues to infer syntactic classes. However, it is by no means trivial, because Catalan syntax is much more flexible than English syntax (Vallduví and Engdahl, 1996) and we use very simple resources, namely, a tagged corpus. If the approach is fruitful, it can be extended to languages with less resources than English or German, such as Catalan itself. The information extracted can be used to create or enhance new resources, such as a parser, and is easy to understand, correct and manipulate by linguists.

In unsupervised techniques, such as clustering, the algorithms do not need any set of pre-classified training instances to compute the solution. Hence, their results are independent from any human classification and depend only on the features and the parameters chosen. It is sensible, therefore, to consider unsupervised techniques to be more empirical than supervised techniques (because the latter do depend on a previous classification).

The paper has the following structure: Section 2 introduces the classification sought; Section 3 explains the materials and methodology (data, features and approach) of the experiment, and Section 4 its results; Section 5 lists further work. Finally, Section 6 presents the conclusions of this paper.

2 Classification

Our initial aim was to distinguish between transitive verbs (those subcategorising for an NP or clausal object), verbs bearing a prepositional object ("prepositional verbs" from now on), and intransitive verbs (without object of any kind). These classes correspond to the most widely cited distinction in both descriptive and theoretical grammar with respect to verbal syntax. However, the first experimental results made us rethink the classification. When computing two clusters, transitive verbs were concentrated in a cluster and intransitive and prepositional verbs in the other one, according to expectations. However, and consistently across experimental settings, when computing more than two clusters, the algorithm made divisions of the transitive cluster, and did not separate intransitive from prepositional verbs.

We believe that this is due to the fact that both intransitives and prepositionals coocur with prepositions and, therefore, they are not different enough to be classified in different clusters. Also, transitive verbs were divided into subclasses because they show a more heterogenous behavior and its number is much greater than the number of both intransitive and prepositional verbs (see in Section 3.1).

As for the divisions within transitives, they were by no means random. A particular class of verbs tended to be separated from more prototypical transitives: Verbs which require an NP object unless they occur with the particle *se*,[1] in which case they require a prepositional object (and admit no NP object), as example 1 shows.[2] We call this class VASE (Verbs Alternating with a *SE*-construction).

(1) a. La revolució no **beneficia** tothom
the revolution not benefits everyone
'Revolution doesn't benefit everyone'

b. L'agricultura es **beneficia** del
the agriculture SE benefits of the
conflicte
conflict
'Agriculture benefits from the conflict'

This class corresponds to an alternation which is very common in Catalan, as well as in other Romance languages (Rosselló, 2002). In our Gold Standard it corresponds to 20% of the lemmata (opposed to 10% intransitive and 8% prepositional; see Section 3.1). Due to the importance of this alternation, and to the fact that these verbs share properties both with transitive and prepositional verbs (they sometimes bear an NP object, sometimes a prepositional one), we found it advisable to add this class to our targeted classification.

In the light of these experimental results, we redefined the classification and designed a two step procedure. In the first step (Sections 3 and 4), the task was to classify verbs into transitive, intransitive and VASE. Intransitives include both verbs subcategorising prepositional objects and pure intransitives. In the second step, briefly explained in Section 5, we further distinguished between prepositional verbs and pure intransitives.

3 Material and method

3.1 Data: Corpus and Gold Standard

We used a 16 million word fragment of the CTILC (*Corpus Informatitzat de la Llengua Catalana*) corpus (Rafel, 1994). The corpus has been automatically annotated and hand-corrected, providing lemma and morphological information (part of speech and inflectional features).

The experiments were carried out on 200 verbs, randomly selected among those having more than 50 occurences in the corpus. To be able to evaluate and analyse the results, one of the authors of the paper classified them into the three classes described in the previous Section. The resulting Gold Standard classification is depicted in Table 1.

	#	%
Transitive	129	64.5
VASE	39	19.5
Intransitive	32	16.0

Table 1: Classes for the Gold Standard.

Note that the largest class is by far that of transitive verbs, and that the intransitive class

[1] *Se* is a morpheme present in the grammar of most Romance languages, which typically absorbs an argument of the verb. There is still debate on whether it absorbs the internal or the external argument. See Bartra (2002) for an overview of its uses in Catalan.

[2] All examples in the paper are taken from the CTILC corpus (see Section 3.1) and shown literally or in an simplified version.

abbrev.	gloss
1 ObjCl	Cooccurence with an object clitic.
2 DetOrN	Determiner or noun follows.
3 Passive	Passive construction.
4 Punct	Punctuation marks (stop, colon, etc.) follow.
5 Prep	Preposition follows (except for *per 'by'*).
6 Se	Particle *se* precedes or follows the verb.
7 DetAndN	DetOrN + precedence by an NP element (adj, pron, det or noun).
8 NonAgrN	DetOrN + not agreement in number.
9 NonAgrP	DetOrN + not agreement in person.
10 NonFin	DetOrN + verb in a nonfinite form.

Table 2: Features used for verb classification.

is the smallest one, despite the fact that it includes verbs bearing a prepositional object and verbs with very unfrequent transitive uses (*dormir la migdiada 'take a nap'*, as transitive use of *dormir 'sleep'*). Taking this distribution into account, we can establish a baseline for the evaluation: Instead of randomly assigning verbs to classes, we will use a higher baseline, that of assigning all verbs to the most common class, transitive verbs. This results in a 0.65 F-score (more details in Section 4).

3.2 Features

We defined ten features suitable to characterise the targeted classes, along with superficial linguistic cues which allowed us to automatically extract the data by simple frequency counts. Table 2 summarizes the features and the shallow cues, and we describe our hypotheses with respect to the characterisation of the classes in what follows.

The first three features, ObjCl, DetOrN and Passive, are directed towards characterising transitive uses of verbs. We expect transitive verbs to have the highest values for these features, while VASE verbs will have middle values but still higher than intransitive ones, due to the uses of VASE verbs where they occur with an NP object.

Note that, as subjects may appear postverbally in Catalan (especially with unaccusative verbs; see sentence (2)), some intransitive verbs may also have relatively high values for feature DetOrN.

(2) **Apareixerà** el monstre
Appear-fut the monster
'The monster will appear'

The following two features, Punct and Prep, are expected to characterise intransitive uses of verbs, so that transitive and (to a lesser extent) VASE verbs are expected to have lower values for them than intransitive verbs.

Feature Se is the only one specifically designed to identify VASE verbs. VASE verbs should have the highest values for this feature and intransitive ones the lowest, since *se* is mostly related to phenomena related to transitivity: reflexivity, passivization, etc.

The last four features, DetAndN, NonAgrN, NonAgrP and NonFin, are aimed specifically at distinguishing transitive verbs from intransitive verbs with a postverbal subject, which is a major problem for our task, as mentioned above and exemplified in sentence (2). The same problem would arise with any other language with a similar syntactic pattern, such as Italian or Spanish. The last features are elaborations on DetOrN designed to detect objects. The restriction that an NP both precedes and follows a verb (feature DetAndN) makes it more likely that an object is present; also, the fact that the NP following the verb does not agree with it in number or person (features NonAgrN and NonAgrP) also point to an object. As for feature NonFin, it exploits the fact that postverbal subjects with infinitives are very rare in Romance languages.

The first six features are represented in terms of raw percentages. Because the last four features are prone to sparse data problems, their values are proportions within the values for DetOrN. The result of the feature extraction is a representation for each verb as in Table 3. We see there e.g. that 9.3% of the occurrences of the verb *contemplar 'contemplate'* (transitive) exhibit the feature ObjCl, while *beneficiar 'benefit'* (VASE) only presents 3% and *xisclar 'scream'* (intransitive) 0%.

Table 4 shows the mean values for each fea-

Lemma	**Class**	**ObjCl**	**DetOrN**	**Passive**	**Punct**	**Prep**
contemplar	Trans.	9.3	52.2	3.4	4.3	15.0
beneficiar	VASE	3.0	20.1	2.5	6.5	32.6
xisclar	Intr.	0	11.7	0	22.0	11.0
		Se	**DetAndN**	**NonAgrN**	**NonAgrP**	**NonFin**
contemplar	Trans.	5.9	15.1	17.3	13.7	25.4
beneficiar	VASE	37.6	39.2	33.3	3.9	19.0
xisclar	Intr.	0.8	0	0	0	6.6

Table 3: Feature values for verbs *contemplar*, *beneficiar*, and *xisclar*.

ture according to the class. [3] Most of the expectations are met: Transitive verbs have the highest values across classes for seven out of the ten features: ObjCl, DetOrN, Passive, DetAndN, NonAgrN, NonAgrP and NonFin. Intransitive verbs have highest values only for Punct and Prep. VASE verbs have intermediate values for most features (the ones for which transitive verbs have high values, plus Prep), high values for Se and low values for Punct. Some of the differences, such as those for Punct, are not as high as expected and may not even be significant, but the patterns are very consistent with our hypotheses.

Feature	**Trans.**	**VASE**	**Intr.**
ObjCl	**4.8**	*4.6*	0.5
DetOrN	**26.4**	*16.3*	14.1
Passive	**6.5**	*3.1*	0.6
Punct	*7.1*	6.8	**10.9**
Prep	17.3	*31.3*	**40.2**
Se	*11.8*	**33.8**	2.6
DetAndN	**31.9**	*27.6*	23.0
NonAgrN	**28.4**	*26.5*	13.2
NonAgrP	**12.4**	*12.2*	3.1
NonFin	**54.6**	*41.7*	18.6

Table 4: Mean values for features by class.

3.3 Clustering approach

We used CLUTO [4] for the experiments. We will report the results obtained with the k-means algorithm. We also tried several of the other algorithms provided with CLUTO (hierarchical and flat, agglomerative and partitional), obtaining quite similar results.

4 Results

With k-means, the number of clusters has to be predetermined. Because our targeted classification consists of three classes, we concentrated on the three cluster solution and will report results for this partition only. As we see in Table 5, cluster 0 contains mainly transitives, cluster 1 intransitives and cluster 2 VASE. Therefore, there is a clear correspondence between classes and clusters, and the cluster analysis has identified the structure we aimed at. However, as detailed in Table 5, there are also some misclassified verbs, which will be further analysed in Section 4.1. Table 6 shows the mean value for each feature in each cluster.

Cluster	Trans.	VASE	Intr.	*Total*
0	**115**	7	5	*127*
1	9	0	**26**	*35*
2	5	**32**	1	*38*
Total	*129*	*39*	*32*	*200*

Table 5: Contingency table.

Feature	**0**	**2**	**1**
ObjCl	**5.2**	*4.0*	0.4
DetOrN	**26.8**	*16.2*	14.0
Passive	**6.7**	*2.9*	1.0
Punct	*7.2*	6.9	**10.2**
Prep	15.2	*33.4*	**44.3**
Se	*10.9*	**38.9**	2.7
DetAndN	**31.2**	*29.2*	24.6
NonAgrN	**29.6**	*26.0*	10.8
NonAgrP	**12.9**	*10.6*	4.3
NonFin	**57.6**	*38.7*	14.1

Table 6: Mean values for every feature for clusters 0, 1 and 2

These data fit with the distribution of feature values across classes reported in Table 4, showing that the value distribution of the features defined for each class is consistent with the predictions. For example, verbs which have middle values for features indicating transitivity tend to have a relatively high value for *Se*.

Table 7 shows the evaluation measures as compared to the Gold Standard: Precision, recall and F-score. As for the baseline, recall

[3] In Tables 4 and 6, the highest mean value appears in bold face, and the middle mean value in italics.

[4] http://www-users.cs.umn.edu/~karypis/cluto/.

from Section 3.1 that we use that of considering all verbs to be transitive, the largest class in the Gold Standard. The overall measures are weighted according to the number of verbs in each class, so that they should be read as the probability of correctly classifying a verb, given the distribution of the Gold Standard across classes.

	Prec.		**Recall**		**F-score**	
Class	Cl.	*(Bl.)*	Cl.	*(Bl.)*	Cl.	*(Bl.)*
Trans.	.91	*(.65)*	.89	*(1)*	.90	*(.82)*
Intr.	.74	*(0)*	.81	*(0)*	.78	*(0)*
VASE	.84	*(0)*	.82	*(0)*	.83	*(0)*
Overall	.87	*(.65)*	.87	*(.65)*	.87	*(.65)*

Table 7: Clustering results (Cl.) compared to baseline *(Bl.)*.

The average F-score is 0.87 a good overall result for a lexical acquisition task, and also compared to the baseline (0.65).

Note that the class with the highest score is that of transitives, probably due to the fact that it is the largest class, and most features are characteristic of transitives, so that the clustering algorithm has richer information for them. Conversely, intransitive verbs get the lowest score. The most plausible explanation, apart from it being the smallest class, is that it contains heterogeneous elements: Pure intransitives and verbs subcategorising for a prepositional object. A second experiment we performed was devoted to that distinction (see Section 5).

4.1 Error analysis

Transitive verbs misclassified into cluster 1-Intr.: *alterar (alter), cessar (dismiss; stop), configurar (set up), consultar (consult), netejar (clean), operar (operate), pensar (think), rectificar (correct), reposar (rest; put again).*

Most of these verbs either are very frequently used without the object (as *netejar* or *operar*) or alternate between an NP and a prepositional object (*cessar de, pensar en*). These verbs are polysemic, and each sense subcategorizes for a different frame. For instance, in the 'stop' sense *cessar* subcategorizes for a prepositional phrase, while in the 'dismiss' sense it is a plain transitive. We didn't establish a specific class for this alternation and therefore classified this verb as transitive. As the 'stop' sense is far more frequent, the feature values for this verb are closer to those of intransitive verbs and, accordingly, it is classified in cluster 1. This second kind of mistake thus points to a richer classification, and the eventual need to encode different frames associated to different senses in case of polysemy. However, this implies a richer lexical representation, which is more difficult to exploit.

Transitive verbs misclassified into cluster 2-VASE: *avorrir (bore), coure (cook), errar (err), espolsar (dust), intensificar (intensify).*

All these verbs appear very frequently with particle *se* in the corpus, most of them due to a causative/noncausative alternation (*El Joan cou la carn 'Joan cooks the meat'* vs. *La carn es cou 'The meat gets cooked/cooks'*). As the noncausative construction is more frequent, they have values similar to VASE verbs. Again, it would be possible to integrate this alternation in the classification, but it affects a comparatively small number of verbs.

Misclassified intransitive verbs: *concordar (agree)* (classified in cluster 2-VASE); *agradar (like), al.ludir (allude), esmorzar (have breakfast), néixer (be born), regalimar (drip)* (classified in cluster 0-Trans.).

Most mistakes in classifying intransitives are due to idiosyncracies of the verbs. For instance, *esmorzar* and *regalimar* have some transitive uses and *agradar* and *néixer* appear almost exclusively with a postverbal subject.

Misclassified VASE verbs: *admirar (admire), afegir (add), aprofitar (make the most), compadir (pity), envoltar (surround), servir (serve; be useful), trobar (find).*

All misclassified VASE verbs are in cluster 0-Trans. These errors are due to the fact that the *se* construction of these verbs (i.e. *admirar-se de, aprofitar-se de*) does not appear often in the corpus, so that these verbs have low values for features Se and Prep and, hence, are more similar to transitive verbs than to VASE verbs.

To sum up, we have seen that the verbs that have been misclassified are in one way or another not **prototypical** within their class. Intuitively, they should also not be similar to the prototype of the class where they have been wrongly placed. A preliminary analysis of the z-scores of the verbs indicate that the intuition is correct for transitive and VASE verbs, but not for intransitive verbs. For two of the clusters, thus, we find that mistakes correspond to distance to the centroid. This suggests that cluster analysis could be used to approach the notion of prototypicality within a class, although further research is needed on this issue.

5 Further work

We are currently testing the system with a 208 million word corpus extracted from the Web (Boleda et al., 2005). With this resource, results are much worse, achieving only a 0.73 F-score (which is however still well beyond the baseline). It is surprising that with on average 12 times the evidence for a verb, results decrease so much; the reason could be the noise contained in such a corpus.

In addition, we have performed another classification experiment which we cannot fully explain due to space constraints. The experiment was aimed at further dividing intransitive verbs into pure intransitives and verbs bearing a prepositional object. The baseline for the task was 0.5 and the upperbound 0.94. Using the experimental setting explained in Section 3 and four features, we achieved an average 0.84 F-score, only 10 points away from the upperbound.

As in the previous experiment, misclassified verbs are verbs whose behavior is closer to the behavior of the verbs of the other class. Most of the missclassified pure intransitives are verbs that very frequently appear with a particular kind of locative adjunct (*conduir per 'drive on', xocar contra 'crash into'*). As for misclassified prepositional verbs, they are those which have some transitive uses (*pujar 'go up, raise', baixar 'go down, lower'*) or that very often appear without the prepositional object (*jugar 'play', protestar 'protest'*).

6 Conclusions

We have presented a cluster analysis which can be used to classify verbs into basic syntactic classes in Catalan using very simple resources (a corpus with morphological information), and which we believe can be straightforwardly extended to other Romance languages, for which there are typically less available resources than for English.

We classified verbs into transitive, intransitive and verbs alternating with a *se*-construction. We defined ten features with their associated shallow cues, which are linguistically motivated and which our experiments have empirically validated. We achieved a mean F-score of 0.87 for an experiment with a 0.65 baseline, which is a good result for a lexical acquisition task.

We have argued that the defined features and the cluster analysis are also useful to determine the prototypicality of a verb within a class. Misclassified verbs are those that have some special property (belong to a subclass, present a particular alternation) and, hence, tend to be further from the centroid of the cluster. Therefore, the mistakes of this system are also linguistically motivated.

7 Acknowledgements

Many thanks to Toni Martí, Enric Vallduví and all the colleagues from the GLiCom for their useful comments. Special thanks are due to the Institut d'Estudis Catalans for lending us the research corpus, and to Nadjet Bouayad and Sebastian Padó for revision of previous versions of this paper. This work is supported by the Departament d'Universitats, Recerca i Societat de la Informació (grants 2003FI-00867 and 2001FI-00582).

References

A. Bartra. 2002. La passiva i les construccions que s'hi relacionen. In J. Solà, editor, *Gramàtica del Català Contemporani*, pages 2111–2179. Empúries, Barcelona.

G. Boleda, S. Bott, B. Poblete, C. Castillo, M.E. Fuenmayor, T. Badia, and V. Lopez. 2005. Cucweb: A catalan corpus built from the web. In preparation.

M. Brent. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 7:243–262.

T. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of ANLP-97*, Washington, USA.

N. Ide and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):1–40.

C. Manning. 1993. Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st ACL*, pages 235–242.

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

J. Rafel. 1994. Un corpus general de referència de la llengua catalana. *Caplletra*, 17:219–250.

J. Rosselló. 2002. El SV, I: verb i arguments verbals. In J. Solà, editor, *Gramàtica del Català Contemporani*, pages 1853–1949. Empúries, Barcelona.

S. Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of COLING-00*, pages 747–753.

S. Stevenson and E. Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of CoNLL-2003*.

E. Vallduví and E. Engdahl. 1996. The Linguistic Rrealization of Information Packaging. *Linguistics*, 34:459–519.

Availability of subcategorization frames: A matter of syntactic or lexical frequency?

Sandra PAPPERT, Johannes SCHLIESSER, & Thomas PECHMANN
Linguistics Department, University of Leipzig
Beethovenstr. 15
04107 Leipzig, Germany
{muckels, schliess, pechmann}@rz.uni-leipzig.de

Dirk P. JANSSEN
Department of Psychology,
University of Kent at Canterbury
Canterbury, Kent CT2 7NP, UK
D.Janssen@kent.ac.uk

Abstract

Psycholinguistic studies investigating syntactic expectations have focussed on the role of linguistic and contextual information. But the availability of subcategorization frames might also be subject to the syntactic frequency of the respective structures or to the lexical frequency of the subcategorizing verbs. We will present data from a completion questionnaire, from a reading time experiment, and from a series of corpus queries that speak to this issue.

1 Introduction

German verb-final sentences are temporally underspecified with respect to argument structure. The aim of the present paper is to evaluate some factors that may affect expectations of verb-specific information before it is available. Recent psycholinguistic studies indicate that processing of verb-final sentences relies on argument-specific information to anticipate the subcategorization frame of the verb (e.g., Friederici and Frisch, 2000; Kamide, Altmann, and Haywood, 2003).

As to the prediction of single vs. double object sentences, factors said to influence word order preferences might also play a role. Empirical evidence for case effects comes from corpus counts (Kempen and Harbusch, 2003), from questionnaire studies (Keller, 2000), and from various reading experiments (e.g., Rösler, Pechmann, Streb, Röder, and Hennighausen, 1998).

In addition to linguistic information as case marking, the frequency of occurrence of a syntactic pattern (Lapata, Keller, and Schulte im Walde, 2001) or that of potentially subcategorizing verbs (Scheepers, Hemforth, and Konieczny, 1999) might modulate the availability of subcategorization frames.

In the following, we present data from a completion questionnaire, from a self-paced reading experiment, and corpus counts to shed light on the impact of syntactic information on argument structure preferences.

2 Completion data

Completion questionnaires allow to investigate the availability of argument structures to subjects. As such, the task involves comprehension processes (to read the fragments) and production processes (to conceive a meaningful continuation).

2.1 Method

A completion questionnaire was assigned to 32 native speakers of German. There were 32 experimental sentence fragments consisting of a subject, an auxiliary, and an object. Nominal constituents referred to animate entities. Case of the object was manipulated.

Der Doktor wird dem/ den Krankenpfleger ...
the$_{NOM}$ doctor will the$_{DAT/ACC}$ nurse ... [1]

The sentence fragments had to be completed by at least a subcategorizing verb to be grammatical. Participants were asked to make the sentences meaningful.

2.2 Results

First, the length of the completions in number of constituents was calculated ignoring the syntactic status of the constituents (see Table 1).

length	proportion
1 constituent	37,4
2 constituents	57,8
3 constituents	4,5

Table 1: Proportion of completions (in %) per number of constituents

Table 1 shows that fragments were most frequently supplemented by two constituents.

[1] *nom* refers to *nominative*, *dat* to *dative* and *acc* to *accusative*.

Second, the proportion of completions including a verb only (single object sentences) or a second object and a verb (double object sentences) were determined for the two conditions (see Table 2).

	dat	acc
single object	13,3	54,5
double object	50,8	3,7

Table 2: Proportion of single vs. double object sentences (in %) per condition

After a dative object, participants tended to insert a second object, whereas after an accusative object, they preferred to complete the fragment with a verb only.

2.3 Discussion

One might expect a "laziness effect" to occur in a completion task, but the length of the resulting sentences indicates that this was not a confound.

By contrast, analyses reveal a strong interaction of (first) object case and subcategorization frame availability. Whereas single object structures were preferred after an accusative, double object structures were prevalent when the given object was marked for dative. This pattern mirrors sentence processing data from Japanese (Kamide, Altmann, and Haywood, 2003).

The linearization preferences for double object sentences found in the questionnaire study are also in accord with linguistic constraints that penalize the assumably scrambled word order with the accusative object preceding the dative object (e.g., Büring, 2001).

3 Self-paced reading data

Data from self-paced reading experiments have been shown to be sensitive to processing difficulties due to syntactic complexity or due to unexpected sentence materials. As reading times are measured segment-by-segment, effects may be located during incremental sentence processing.

3.1 Method

36 native speakers of German participated in the experiment. The experimental set consisted of 32 double object sentences with the subject in first position and the subcategorizing verb in final position. Referents of the nominal constituents were animate. The order of the objects was manipulated:

> Der Doktor wird dem/den Krankenpfleger den/dem Rollstuhlfahrer zeigen.
>
> the$_{\mathrm{NOM}}$ doctor will the$_{\mathrm{DAT/ACC}}$ nurse the$_{\mathrm{ACC/DAT}}$ wheelchair person point out to
>
> 'The doctor will point the nurse/wheelchair person out to the wheelchair person/nurse.'

To prevent participants from predicting the sentences' length, 32 filler sentences with two nominal constituents were added to the list.

Sentences were presented word-by-word, with no hint to length. Each sentence was followed by a case-sensitive content question.

> Wird der Doktor den Rollstuhlfahrer dem Krankenpfleger zeigen?
>
> 'Will the doctor point the wheelchair person out to the nurse?'

3.2 Results

First, errors in answering the content question were computed. A high error rate (mean: 26 %) was found, but there was no significant effect of word order. Error trials were excluded from further analysis.

Reading time analyses revealed no significant differences on words 1 to 5. But on word 6, the determiner of the second object (*den* vs. *dem* [*Rollstuhlfahrer*], *the$_{ACC}$* vs. *the$_{DAT}$* [*wheelchair person*]), a significant effect of word order was found (t_1 (35) = 2.95, $p < .01$; t_2 (30) = 2.08, $p < .05$). Reading times were longer in the nom-acc-dat condition than in the nom-dat-acc condition (see Table 3).

nom-dat-acc	nom-acc-dat
571	611

Table 3: Mean reading times (in ms) on word 6 (*den/dem*) per condition

On the following words 7 and 8, no significant differences arose.

3.3 Discussion

The high error rate indicates that subjects experienced problems while processing the case information, be it during reading of the experimental sentences or during answering of the content questions.[2]

As there was no significant difference in reading times on the first object, a specific problem with

[2] We will redo the experiment with sentence matching as additional task. Case will be varied, but noun order will be kept constant. Accordingly, we expect the additional task to produce less errors than the answering of content questions.

the processing a dative vs. an accusative object can be excluded.

By contrast, a word order effect was found on the determiner of the second object. Given the single object sentences that were included as fillers, the effect occurred as soon as the local indeterminacy concerning the argument structure was resolved in favour of a double object reading. We interpret this effect as evidence of a double object expectation after dative objects and a single object expectation after accusative objects.

4 Syntactic frequency data

To ascertain whether the frequency of occurence of single and double object sentences can help to predict the performance in the completion questionnaire and in the self-paced reading experiment, corpus counts were carried out.

4.1 Method

From Negra2 and Tiger, two syntactically annotated newspaper corpora, single and double object sentences were extracted with the nominative constituent topicalized and the subcategorizing verb in final position.

4.2 Results

There were 4737 sentences that met the above mentioned criteria and that did not contain pronouns (see Table 4).

	nom-dat ...	nom-acc ...
2 arguments	336	4205
3 arguments	176	20

Table 4: Number of syntactic structures per subcategorization frame in Negra2 and Tiger

In this corpus subset, single object sentences are much more frequent than double object sentences. This difference is especially huge for the sentences with the (first) object marked for accusative.

4.3 Discussion

The data match those reported by Kempen and Harbusch (2003). On the basis of the corpus data, one would predict that single object sentences were preferred over double object sentences and that this preference was even more pronounced when the (first) object is marked for accusative.

However, this prediction was not met by the completion data. Indeed, there was no overall preference of single object sentences, but the availability of different argument structures hinged on the case marking of the given object. By consequence, syntactic frequencies are disqualified as a predictor of completion performance.

5 Lexical frequency data

Alternatively, the lexical frequency of the potentially subcategorizing verbs might help to predict the availability of syntactic frames in a completion questionnaire.

5.1 Method

The German Syntax part of the Celex corpus (Baayen, Piepenbrock, and L. Gulikers, 1995) provides information about lexical frequency as well as about obligatory and impossible verb complements.

5.2 Results

The corpus includes 7232 verbs that must *obligatorily* take a dative and/or an accusative complement and 7738 verbs that can *potentially* do so.

First, absolute frequencies for the different verb types were summed (see Table 5).

	obligatory	potential
dat	220713	199129
acc	481729	425055
dat&acc	90891	206521

Table 5: Summed absolute frequencies of verb types with an obligatory vs. potential subcategorization frame in Celex

Verbs that (obligatorily or potentially) occur with an accusative single object have the highest frequency of occurrence. As for obligatory subcategorization frames, verbs that take a single dative object rank in frequency above those that take two objects. For potential subcategorization frames, these the summed frequencies of these two verb types do not differ.

Second, the number of different verb tokens per obligatory and potential subcategorization frame was computed (see Table 6).

	obligatory	potential
dat	234	225
acc	6336	6294
dat&acc	662	1219

Table 6: Number of verb tokens per obligatory vs. potential subcategorization frame in Celex

Counts on verb tokens reveal a strong prevalence of verbs subcategorizing a single accusative object over. In addition, there are less verb tokens subcategorizing a single dative object than tokens subcategorizing two objects.

5.3 Discussion

As the syntactic frequency of the syntactic frames, the summed frequency of the verb types does not predict performance in the completion questionnaire and in the reading experiment. By contrast, the number of verb tokens with a specific subcategorization frame might modulate the availability of the respective frames. However, as Celex does not report data on word order in double object sentences, frequency and number counts can not be attributed to double object sentences with the one vs. the other order of objects.

6 Syntactic frequency data reconsidered

As syntactic corpora provide information about word order variation in double object sentences, a reconsideration seems worthwhile. There is one potential confound in the syntactic frequency data reported above that might be ruled out: Whereas animacy was controlled for in the completion questionnaire and in the self-paced reading experiment, it was not considered in the corpus query.

6.1 Method

The 4737 sentences from the Negra2 and Tiger subset that matched the syntactic structures of the experimental sentences were manually annotated for animacy. There were three categories, one including clearly animates as humans and animals, one including intermediate entities as institutions, artefacts acting as humans (e.g., cars) etc., and one including clearly inanimates as non-acting artefacts. A conservative count excluded members of the intermediate category, a more permissive count recognized them as animates.

6.2 Results

Only counts on sentences with animate referents of the subject and the (first) object will be reported, one excluding (see Table 7) and one including the intermediate category (see Table 8).

	nom-dat ...	nom-acc ...
2 arguments	29	141
3 arguments	45	0

Table 7: Number of syntactic structures per subcategorization frame in Negra2 and Tiger, only subjects and (first) objects referring to humans and animals included

	nom-dat ...	nom-acc ...
2 arguments	85	452
3 arguments	130	0

Table 8: Number of syntactic structures per subcategorization frame in Negra2 and Tiger, also subjects and (first) objects referring to institutions etc. included

The conservative and the permissive counts show a similar pattern: Whereas sentences with a single accusative object are more frequent than sentences with an accusative preceding a dative object, sentences with a dative preceding an accusative object are relatively more frequently than single dative object sentences.

6.3 Discussion

Syntactic corpus data finally account for the behavioural data when animacy as a potential confound is taken into consideration. Counts excluding inanimate referents reveal a prevalence of single object sentences with the object marked for accusative as well as a (weaker) prevalence of double objects when the (first) object is marked for dative.

As datives and accusatives clearly pattern differently, animacy of the constituents' referents alone could not account for the behavioural data.

To conclude, neither syntactic nor semantic information alone modulates the availability of subcategorization frames in a completion or a reading task, but the interaction of both is crucial.

7 General discussion

To summarize, the completion data and the reading time data on single vs. double object sentences are not accounted for by syntactic frequency counts that ignore semantic information. A lexical measure, the summed frequency of the verb frame types does not predict performance in the completion task either. But another lexical measure, the number of verb tokens may function as a predictor of the availability of the respective syntactic frames. And finally, corpus counts that take syntactic and semantic information into consideration may account for the behavioural data.

The data reported here indicate that processing of verb-final sentences may profit from a rich evaluation of argument-specific information as case and animacy. In future experiments, animacy information carried by the nominal constituents will be manipulated. These experiments will help to settle the issue of how syntactic and semantic information interact during incremental sentence processing.

As a first consequence of our results, we postulate that predictions of subjects' performance in psycholinguistic experiments should generally be based on corpus data that give information about the syntactic *and* the semantic distribution of the items.

8 Acknowledgements

The work presented here emerged from the project "Predictive sentence processing in German" which is part of the research program "Linguistic foundations of cognitive science" funded by the Deutsche Forschungsgemeinschaft (DFG), see http://www.uni-leipzig.de/~parsing/.

References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. *The CELEX Lexical Database* [CD-ROM]. Linguistic Data Consortium, Philadelphia, PA.

D. Büring. 2001. *Let's phrase it! Focus, word order, and prosodic phrasing in German double object constructions.* In "Competition in syntax", G. Müller & W. Sternefeld, ed., pages 69-105, DeGruyter, Berlin.

A. D. Friederici and S. Frisch. 2000. Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and Language,* 43: 476-507.

Y. Kamide, G. T. M. Altmann and S. L. Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language,* 49: 133-146.F. Keller. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD thesis, University of Edinburgh.

G. Kempen and K. Harbusch. 2003. An artificial opposition between grammaticality and frequency: Comment on Bornkessel, Schlesewsky & Friederici (2002). *Cognition,* 90: 205-210.

M. Lapata, F. Keller, and S. Schulte im Walde. 2001. Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research,* 30: 419-435.

[Negra2] http://www.coli.uni-sb.de/sfb378/negra-corpus/F. Rösler, T. Pechmann, J. Streb, B. Röder, and E. Hennighausen. 1998. Parsing of sentences in a language with varying word order: Word-by-word variations of processing demands are revealed by event-related brain potentials. *Journal of Memory and Language,* 38: 150–176.

C. Scheepers, B. Hemforth, and L. Konieczny. 1999. *Incremental processing of verb-final constructions: Predicting the verb's minimum (!) valency.* Paper, International Conference on Cognitive Science, Tokyo.

[Tiger] http://www.ims.uni-stuttgart.de/projekte/TIGER/

Postulation of Lexical Semantic Types for a Set of Intransitive Verbs in Bangla

Soma Paul
Language Technology Research Centre
International Institute of Information Technology
Hyderabad 500019
India
soma@iiit.net

Abstract

This paper examines the semantic feature structure of a set of intransitive verbs that takes as subject an argument that entails to undergo a change of state (physical or mental). While all these verbs are defined to be the subtype of a general semantic type (I will call it *undergoer type*), I propose to incorporate additional semantic features into their semantic feature structure that classify the semantics of these verbs into different subtypes. This paper contends that differences in semantic structure will account for their syntactic behavioral differences. Two verbs combine to form a verbal complex in case they are semantically compatible.

1 Introduction

There are varied opinions and wide-ranging speculation regarding what and how much should be incorporated into the semantic representation of a verb. I take the popular approach[1], which sets the goal to identify those aspects of lexical semantics that are "grammatically relevant"; in other words, the semantic factors that affect the syntactic behavior of a verb and account for the range of alternation the verb undergoes.

I will discuss here about a set of monadic verbs whose argument that occupies the subject position satisfies one of the following entailments:

1. Undergoes instantaneous change of state in event.
2. Undergoes gradual (incremental or decremental) change of state in event
3. Undergoes a change of state and entails to be located at a resultant state that follows the change.

This paper attempts to postulate a set of *semantic types* each of which designates an event in which the participant undergoing a change of state bears one of the above entailments. The paper contends that semantic structure of the verbs under consideration accounts for the differences in their syntactic behavior.

I adopt Davis's (2001) model of lexical semantic representation. The semantic content of a verb is represented as *typed feature structure*[2]. The value of each proto-role attribute within the *semantic type* denotes an entity that plays a certain participant role in the denoted situation. Playing that role implies that a proto-role entailment associated with that attribute will hold of that entity by virtue of its participation in the situation. The idea of associating classes of entailments with proto role attributes follows Dowty's (1991) proposal of proto-role model, which relies on a set of entailments to determine the mapping between semantic role and syntactic arguments. For example, the *typed feature structure* with the type designation *undergoer type* in figure 1 constitutes the linguistic representation of an event that corresponds to a situation in which a participant is entailed to undergo one of the following:

[1] Pinker (1989), Jackendoff (1990), Levin (1993), Levin and Rappaport (1995, 1998) and Wechsler (1995) work in this direction.

[2] Carpenter (1992) made an in-depth treatment of these formal foundations.

a. To undergo a change of state
b. To undergo movement
c. To be causally affected
d. To be located[3].

undergoer type
UND

Figure: 1

The semantic information about the lexical entailments is given in this paper within a *configurational structure* rather than simply listing them on a single plane. These representations bear some resemblance to the decomposition model of Jackendoff (1989) and Pinker (1990). Thus an argument may bear one proto-role in one subevent and a different one in another.

The semantic types along with the list of verbs whose semantic content correspond to these types are specified in section 2. Section 3 records how the syntactic behavior of verbs of one class varies from those of the other. Section 4 proposes a mechanism that accounts for the distinction of the syntactic behavior of verbs under consideration.

2 Semantic types and their formal representation

In this paper I identify three broad semantic types. They all are the subtypes of the general type *undergoer type* (as discussed in section 1). This type contains only one attribute UND as shown in figure 1.

2.1 Change of state type

The *change of state type* corresponds to a situation in which a participant is entailed to undergo a change of state. This relation type subsumes the semantics type of the following verbs:

Class1:

bhaŋa 'break' *chẽṛa* 'tear' *khola* 'open'
khɔša 'fall off' *oba* 'evaporate' *mɔra* 'die'
phaṭa 'explode' *thæ̃tlano* 'get smashed'
olṭano 'tumble' *mɔckano* 'get twisted'
ghoca 'disappear' *opcano* 'spill'

These verbs represent the function described by the operator BECOME in Dowty's work. It is defined as

$$\text{BECOME}(p) =_{\text{def}} \sim pTp,$$

where p is a state, T is a dyadic operator meaning "And Next".

The semantic representation of the *change of state type* is given in figure 2:

change of state type
UND undergoing a change of state

Figure: 2

2.2 Incremental change type

There is another kind of change of state verb that denotes an event type in which the participant undergoes an incremental change (Dowty (1991) first used the term *incremental theme*):

Class2:

kɔma 'diminish' *phurono* 'get exhausted'
ḍoba 'sink' *gɔla* 'melt' *pɔca* 'rot'
poṛa 'burn' *nebha* 'be extinguished' '

The event denoted by verbs of class (2), unlike that denoted by those of class (1), characterizes an internal development. The participant involved in the event undergoes a gradual change. Dowty's *degree achievement* predicates (Dowty 1979) express similar characteristics.

The *incremental change relation* is presented in figure 3:

[3] Within the framework of Role and Reference grammar (1984) two semantic macro-roles, actor and undergoer, have been postulated. As Van Valin notes (2001, p. 30-31), each of these macro-roles represents a grouping of thematic relations mainly for the purpose of defining generalized linking constraints. Along with other thematic relations such as patient, experiencer, recipient the theme (a cluster of semantic roles such as thing located, thing moved, thing given and so on) is represented as undergoer.

$$\begin{bmatrix} \textit{incremental change type} & \\ \text{UND} & \text{incremental theme}^{4} \end{bmatrix}$$

Figure: 3

2.3 Inchoative type

Inchoative verbs denote event types in which the participant that undergoes a change of state is also entailed to be located at a resultant state that follows the change. Therefore the semantic structure of these verbs embeds a subevent of the type *resultant state relation* as shown in the following figure:

$$\begin{bmatrix} \textit{inchoative type} & \\ \text{UND} & \text{[1] undergoing change of state} \\ \text{SOA} & \begin{bmatrix} \textit{resultant state type} & \\ \text{UND} & \text{[1] being located} \end{bmatrix} \end{bmatrix}$$

Figure: 4

The semantic types of the following verbs are the subtype of *inchoative type*:

Class 3a:

ghumono 'sleep' *jhimono* 'doze' *paka* 'ripen'
šoo̯a 'lie down' *phola* 'swell' *jɔla* 'blaze, shine'
phoṭa 'blossom' *bãca* 'survive'

For instance, the verb *phoṭa* 'blossom' denotes an event type in which the participant is entailed both to undergo a change of state, from being 'not blossomed' to being 'blossomed', and to remain in that blossomed state. The embedded subevent denotes the resultant state.

2.3.1 Inchoative incr type

Like verbs of class (3a) the semantics of verbs of class (3b) also implies that the participant undergoes a change of state. However the change is gradual and not instantaneous.

Class 3b:

thama 'stop' *thitono* 'become quite'
bhɔra 'fill'

$$\begin{bmatrix} \textit{inchoative incr type} & \\ \text{UND} & \text{[1] incremental(or decremental) theme} \\ \text{SOA}^{5} & \begin{bmatrix} \textit{resultant state type} & \\ \text{UND} & \text{[1]} \end{bmatrix} \end{bmatrix}$$

Figure: 5

For instance, the sentence in (1a) suggests that *brišṭi* 'rain' has not stopped entirely but it is slowing down gradually:

1a. *brišṭi* ***them-e aš-che***
rain stop-cp come-3 pr cont
'The rain is slowing down gradually /
The rain is about to stop'

The event types denoted by verbs of class (3a), however, do not indicate that the change is gradual. Therefore the following sentence is bad:

1b. **bacca-ṭa* ***šu-e aš-che***
child-cl lie down-cp come-3 pr cont
'The child is about to lie down'

The following hierarchy presents a network of the different semantic types, which have been identified in this section:

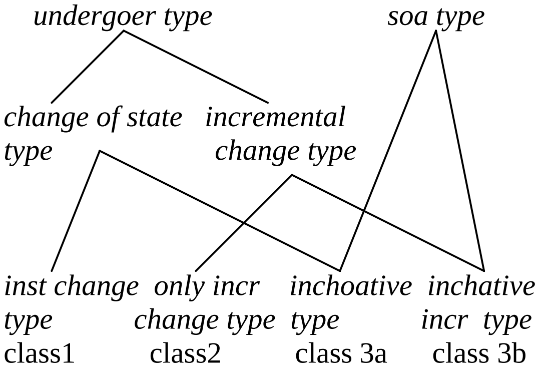


Figure: 6

[4] The labels *theme* and so on are purely mnemonic. The interpretation of the proto-roles is determined by the type that contains the proto-role.

[5] The value of the proto-role attribute SOA in my system is always a subevent that either accompanies the main event or is resulted from the main event.

The next section presents a few contexts where the syntactic behavior of verbs of classes 1, 2 and 3 vary.

3 Differences in syntactic behavior

I will examine two contexts in which verbs under consideration require different treatment. In Bangla[6] simple verbs combine with light verbs as well as auxiliary verbs and the verbal complex[7] are constructed. *aša* 'come' is a light verb that combines with the conjunctive participial of main verb and compound verb sequences[8] are formed. The auxiliary verb *ach* 'be' also combines with conjunctive participial form of verbs. However neither the light verb nor the auxiliary combines with all main verbs. For instance, verbs of class (2) and (3b) occurs with the light verb *aša* 'come' as exemplified in the following sentences:

2a. *alo* ***kom-e aš-e***
light fade-cp come-3 pr
'The light (gradually) fades'

b. *šurjo pošcim akaš-e* ***ḍub-e aš-e***
sun west sky-loc set-cp come-3 pr
'The sun is setting (slowly) in the western sky'

c. *akaš megh-e* ***ḍhek-e aš-e***
sky cloud-loc cover-cp come-3 pr
'The sky is almost overcast'

The V1 participants *kɔma* 'reduce', (*šurjo*) *ḍoba* '(sun) set' and *ḍhaka* 'cover' in the sentences in (2)a, b and c respectively entails a change of state. The change is not, however, instantaneous. On the contrary the event denoted by these verbs involves stages through which the event progresses towards the culmination point, which indicates the change of state. When these verbs select the light verb *aša* 'come', only the developmental stages are profiled. For instance, the compound verb *ḍube aša* 'set (gradually)' does not entail that the sun has already set; it implies that the sun is gradually setting. Thus the event denoted by the verb sequence *ḍube aša* 'set (gradually)' is *atelic* in nature. My presumption is the following: the culmination point of the V1 event is the stationary reference point. The light verb *aša* 'come' focuses on the preliminary stages of the event. Thus the event represented by the simple verb + *aša* 'come' implies directedness towards the culmination point of the V1 event.

The event types denoted by verbs of class (3b) and (1), however, do not entail that the change is gradual. Therefore the following sentence is bad:

2d. **bacca-ṭa* ***šu-e aš-che***
child-cl lie down-cp come-3 pr cont
'The child is about to lie down'

The verbs of *inchoative type* (exemplified in class 3) occur with the verb *ach* 'be' and the verb complex entails that the resultant state that results from the change of state prevails. Thus when these verbs occur in the context of the stative verb *ach* 'be' the stative segment of the event is focused. Verbs of class (3) are compatible with *ach* 'be' as illustrated in the sentences in (3) in contrast with those of classes (1) and (2) (as exemplified in (4)):

3a. *šara akaš megh-e* ***ḍhek-e ach-e***
whole sky cloud-loc cover-cp be-3 pr
'The whole sky is covered with clouds'

b. *tar du-cokh jɔl-e* **bhor-e ach-e**
he-gen two-eye water-loc fill-cp be-3 pr
'Her eyes are filled with tears'

c. *gaṛi-ṭa* ***them-e ach-e***
car-cl stop-cp be-3 pr
'The car is standing (still)'

4a. **šurjo pošcim akaše* ***ḍub-e ach-e***
sun west sky-loc set-cp be-3 pr
'The sun sets in the western sky'

[6] Bangla (popularly known as Bengali) is an Indo-Aryan language spoken in Bangladesh and at Eastern Zone in India.

[7] In Paul (2004) I have proposed that in Indo-Aryan languages predicates (a functional-semantic unit) can be expressed both synthetically (by one word expressions) and analytically (by multi-word expressions). This implies that there is no one-to-one mapping between the meaning-form and physical form of expressions. Multi-word expressions that are composed of more than one grammatical element (either morphemes or words), each of them contributing part of the information ordinarily associated with a head, are usually referred to as complex predicate constructions in modern day parlance.

[8] A compound verb (CV) construction in Indo-Aryan languages is popularly characterized as a kind of complex predicate.

b.*_tak theke poṛ-e glaš-ṭa_ _**bheŋ-e ach-e**_
shelf from fall-cp glass-cl break-cp be-3pr
‘The glass has fallen down from the shelf
and remained in a broken state’

c. *_rou̯dr-e jamakapoṛ_ _**šuki-e ach-e**_
sun-loc clothes dry-cp be-3 pr
‘The clothes dry in the sun’

This paper proposes that the composition of verbal complex is largely determined at the level of semantics because the two verbs will unify if and only if they are semantically compatible. The next section describes the mechanism.

4 Semantically compatible verbs unify

The stipulation of semantic compatibility requires that the semantic entailments of proto-role attributes within the semantic type of the main verb and the light verb or the auxiliary must be compatible or consistent. For instance the verb *ghumono* ‘sleep’ and the light verb *aša* ‘come’ are not semantically compatible. The verb *ghumono* ‘sleep’ denotes an event type in which a participant is entailed both to undergo a change of state (the change is not gradual) and to be located in a state that follows the change (i.e., the state of sleeping). Thus the semantic type of the verb *ghumono* ‘sleep’ will be a subtype of the *inchoative type*. The event of *ghumono* ‘sleep’ as the import of the verb implies does not include the process of getting asleep. The semantic type of the light verb *aša* ‘come’, on the other hand, will be a subtype of *incremental change type.* The full verb counterpart of the light verb *aša* ‘come’ implies *the directedness of a participant towards a stationary reference point* as exemplified in (5):

5a. _ritu amar dike_ _**e-lo**_
Ritu I-gen towards come-3 pt
‘Ritu came towards me’

b. _puronodin-er kɔtha tar mon-e_ _**ašche**_
old days-gen word his mind-loc coming
‘Memories of old days are surging back in
his mind’

ami ‘I’ in (5a) and *mon* ‘mind’ in (5b) are the stationary reference points towards which the other participants is directed. If the grammar licenses the semantic types representing the meaning of the verbs *ghumono* ‘sleep’ and *aša* ‘come’ to unify, the resultant semantic type will correspond to a situation in which a participant will be entailed to be both located and approaching towards a culmination point at the same time. Certainly such a semantic interpretation is ill-formed. Therefore the semantic types of the verbs *ghumono* ‘sleep’ and *aša* ‘come’ are declared inconsistent in the grammar in order to ensure that the compound verb *ghumie aša* “approaching to sleep” is not licensed by the grammar. The semantic type of *ach* is a subtype of *state relation*. I adapt Pinker’s notion of stativity in my grammar by postulating a subtype of *undergoer type*. I call it *stative type*. The value of UND in *stative type* denotes a participant that is entailed to ‘be located’. Since all inchoative verbs have an embedded subevent that denotes stative eventuality, the semantic type of *ghumono* and for that matter any other verb of *inchoative type* is compatible with the semantic type of the verb *ach* ‘be’.

The semantic types that constitute the meaning component of verbs are arranged in a multiple inheritance hierarchy network as shown in figure (7). No two inconsistent types will have a unique greatest lower bound, i.e., a common subtype specified in the hierarchy.

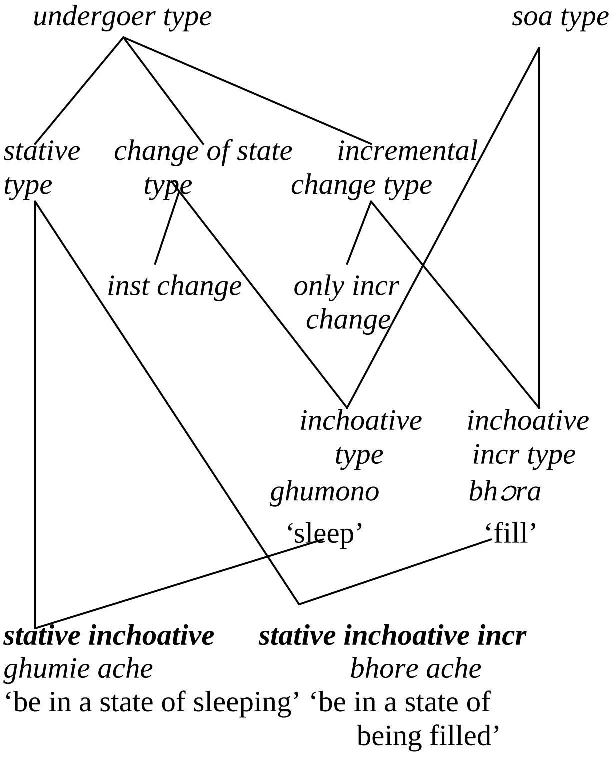


Figure: 7

5 Conclusion

This paper postulates semantic feature structure for a class of verbs that denote an event in which the participant undergoes a change of state. The semantic types representing the semantic content of verbs discussed in this paper are arranged in a multiple inheritance network. I have proposed that these verbs can occur in the context of light verb *aša* ‘come’ and auxiliary such as *ach* ‘be’ only when they semantically compatibility. By semantic compatibility I understand here that the semantic type of one verb is the subtype of the other or they have a common subtype.

6 Acknowledgements

My thanks go to my supervisor Professor Gautam Sengupta and co-supervisor Professor Probal Dasgupta for their guidance and useful suggestions in writing this paper.

References

Anthony R. Davis. 2001. *Linking by Types in the Hierarchical Lexicon.* CSLI publications, Stanford, CA.

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

B. R Levin and Rappaport H. M. 1995. *Unaccusativity, At the Syntax-Lexical Semantics Interface*. Cambridge, Mass.: MIT Press.

David Dowty. 1979. *Word meaning and Montague grammar.* Clarendon Press, Oxford.

David Dowty. 1991. “Thematic Proto-roles and argument selection.” *Language*, 67:3, 547-619.

Hovav Rappaport and Beth Levin. 1998. “Building verb meanings”. In Miriam Butt and W. Geuder (eds.) *The Projection of Arguments*: *Lexical and Compositional Factors*. CSLI Publications, Stanford.

M. Butt. 1995. *The Structure of Complex Predicates in Urdu.* Doctoral Dissertation, Stanford University.

R. Carpenter. 1992. *The Logic of Typed Feature Structures,* (Tracts in Theoretical Computer Science). Cambridge University Press, Cambridge.

R. Jackendoff. 1990. *Semantic Structures.* MIT Press, Cambridge, MA.

Robert D. Van Valin, Jr. 2001. *An Introduction to Syntax*. Cambridge University Press.

Steven Pinker. 1989. *Learnability and Cognition.* MIT Press, Cambridge MA.

S. Wechsler. 1995. *The Semantic Basis of Argument Structure.* CSLI Publications, Stanford, CA.

Verb-instrument information during on-line processing

Rachel Shirley SUSSMAN
Dept of Linguistics
University of Rochester
Rochester, NY 14627 USA
rss@ling.rochester.edu

Ellen CAMPANA
Dept of Brain and Cognitive Science
University of Rochester
Rochester, NY 14627 USA
ecampana@bcs.rochester.edu

Michael K. TANENHAUS
Dept of Brain and Cognitive Science
University of Rochester
Rochester, NY 14627 USA
mtan@bcs.rochester.edu

Greg CARLSON
Dept of Linguistics
University of Rochester
Rochester, NY 14627 USA
carlson@ling.rochester.edu

Abstract

The results of three eye-tracking experiments provide support for a model of on-line processing that includes immediate access to verb representations, including information about instruments likely to be used in the denoted action.

1 Introduction: verb information during sentence processing

Over the past twenty years, research has yielded a growing body of evidence to support the immediate influence of verbs and their associated information on the course of sentence processing. Work generating this evidence has progressed along three major lines: 1) reading tasks to showing that in cases of syntactic ambiguity, the syntactic preferences of the specific verb used in the construction will influence the way in which upcoming material is parsed (e.g. Trueswell, Tanenhaus and Kello (1993), Garnsey, Pearlmutter, Myers, and Lotocky, (1998), and Hare, McRae and Elman (in press), inter alia); 2) reading tasks focused on filler-gap constructions such as questions and relative clauses (e.g. Boland, Tanenhaus, Garnsey, and Carlson (1995)); and 3) analysis of anticipatory eye-movements during the processing of unambiguous, declarative constructions (e.g. Altmann and Kamide (1999), Boland (2002))

Overall, these studies provide ample evidence to support the case for verb-based information playing an important role during processing. However, a certain weakness remains in that the bulk of this evidence comes from studies focusing on how verbs can influence listener/reader predictions about immediately upcoming constituents. This fact creates a potential confound in the interpretation of the verb-based effects seen in the experiments. In particular, due to the sequential/incremental nature of sentence processing, it is already fully expected that a large share of processing resources will be devoted to predicting and integrating the immediate upcoming input. Work from the statistical learning community has independently established that upon being presented with strings of “language” made up of nonsense syllables, listeners unconsciously perform complex calculations of the transitional probabilities between co-occurring syllables. They can then use this knowledge to make inferences about what groupings are likely to be grammatical sequences in the nonsense language (Saffran, Aslin, and Newport 1996). Extending this finding to the realm of verbs, we might expect that after a lifetime of exposure to strings of English and English verbs, the average speaker could easily be expected to have insight about the sorts of elements that are likely to follow the verb in any given sentence. That is, independent of any information that might be part of a verb's linguistic representation, it may be the case that the long-term calculation of co-occurrence information typical to language has given the speaker access to information about syntactic categories, general semantic features, and even particular words that are likely to follow the verb. With this possibility unaccounted for, the effects seen in many of the experiments investigating the role of verb-based information may only hold in cases where predictions can be made about immediately upcoming constituents. While it might be easy to show that verb information is accessed and used to make predictions about what is likely to come next, to truly make the case that the full range of information associated with a verb is integrated into the unfolding interpretation, we would need to see evidence of predictions being made about non-consecutive elements of a verb's argument structure and conceptual information.

The general emphasis on adjacent constituents during sentence processing gives rise to another potential confound, this time concerning systematic differences in the relative strength of effects between adjacent and non-adjacent constituents. Since it is already expected that a large share of processing resources will be devoted to immediate upcoming constituents, activation for these elements is predicted to be larger than for downstream arguments. Therefore, even in cases where evidence for downstream arguments might be observed at the verb, it will likely be overshaddowed by predictions that are being made about immediate upcoming arguments. Results presented in Boland 2002 may reflect such a result. In this study, Boland observed anticipatory looks to recipient arguments at the verb, but not to instruments. This result was attributed to the difference in argument status between recipients (core argument of the verb) and instruments (adjunct to the verb). In addition to argument status however, recipients and instruments also differ in terms of the surface positions they may occupy within a sentence. Unlike instruments, recipients can immediately follow the verb when they occur in dative shifted constructions. In sentence 3a, the recipient "Mary" occurs in a prepositional phrase somewhat downstream from the verb. In the dative shifted version of the sentence shown in 3b, however, the recipient directly follows the verb.

1a. Fred gave an umbrella to Mary.
1b. Fred gave Mary an umbrella.

Thus, in the Boland experiment, the increased looks to recipient arguments relative to instruments may be the result of a certain percentage of participants anticipating the dative construction, and accordingly devoting the majority of their processing resources towards predictions about the immediately upcoming recipient. Since there is no analogous "instrument shifted" construction, there would be no corresponding percentage of participants anticipating an instrument directly following instrument verbs, which in turn would result in a smaller amount of anticipatory looks to instruments. Thus, the observed difference between anticipatory looks to recipients and anticipatory looks to instruments may be an effect of position within the construction rather than of argument status; even if access to the verb provides access to the instrument role, due to its downstream position in the sentence, we expect that roles that may occur adjacent to the verb (such as recipients) will receive the larger share of activation.

2 Background: Verbs and Instruments

Instruments have long been of interest to researchers because of their intermediate status with respect to verbs. While instruments are generally thought to bear a strong semantic connection to the verbs they are associated with, often being an essential component of the action the verb describes, syntactically speaking they are somewhat removed from the verb, appearing in optional adjunct phrases if they appear at all. Indeed, there is evidence to suggest that contrary to findings regarding other elements associated with verbs, a close semantic connection between verb and instrument results in a *decreased* expectation that that instrument will be mentioned in the overt syntax of the upcoming material (Kear and Wilson, 2000). In this way, verb-induced expectations about instruments can provide an ideal testing ground with regard to critical questions of verb based access to syntactic vs. non-syntactic material.

The earliest studies looking for activation of instruments associated with verbs tended to focus on activation of instruments in the context of inferences calculated from the meaning of the sentence as a whole rather than part of a verb's specific lexical information. Generally, these experiments would present participants with an initial complete sentence. The sentence would either overtly mention an instrument, or heavily imply that some sort of instrument must have been used, given the nature of the action described. The second part of the task varied from experiment to experiment, but was generally designed to test the degree of activation for the instrument. It was hypothesized that if participants are routinely calculating instrument inferences, sentences that overtly mention the instrument and sentences that strongly imply the use of an instrument but do not mention it overtly should show the same pattern of results. One experiment of this type found precisely this result (Garrod and Sanford, 1981). However, a large number of similar experiments showed the opposite. In these cases, activation for the instrument was stronger when it had been overtly mentioned earlier in the experiment. This was taken as evidence that inferences about instruments are not automatically calculated, even when the meaning of the sentence strongly implies that an instrument must have been involved in the action (e.g. Singer, 1979, McKoon and Ratcliff, 1981, Dosher and Corbett, 1983, Lucas, Tanenhaus and Carlson 1990).

More recent studies have characterized the relationship between verbs and their instruments as one of meaning rather than inference, and make the assumption that instrument effects will be more closely tied to lexical activation rather than cross sentential reasoning. Ferretti, McRae, and Hatherell (2001) show that in single word priming tasks, verbs such as “stirred” reliably prime their instruments (here, “spoon”). This result paralleled priming between verb and agent (“sketching” – “artist”) and verb and patient (“adopt” – “baby”). By contrast, they found that verbs do not prime their locations; there was no advantage in lexical decision tasks for words like “kitchen” when primed by closely related verbs (i.e., “cooking”). Based on this evidence, the authors conclude that whatever the relationship between verbs and their instruments, it is more akin to the relationship between verbs and their agents and patients than it is to the connection between verbs and their likely locations; that is, whatever representational status is afforded to agents and patients should also be true of instruments.

In an eye-tracking study, Boland (2002) showed that upon encountering a verb with a dative argument structure, participants were more likely to look at a potential recipient argument than they were to look at potential instruments in conditions where they had been presented with a verb that necessitated instrument use. Boland argued that this is a result of the differing argument status associated with recipients and instruments. While recipients are uncontroversially considered to be stipulated as part of the argument structure of any dative verb, the argument status of instruments is less clear. Though like arguments, instruments often bear a strong semantic relationship to verbs, syntactically speaking they pattern more like adjuncts. Boland specifically chose verbs that maintained that strong semantic connection between verb and instrument at the same time that the instrument was unambiguously a syntactic adjunct. On these grounds, Boland interprets her result as evidence that verb argument structure rather than semantic association is the primary source of specific lexical information introduced by the verb. However, due to the confounding of questions of argument status and questions of sequence (discussed above) it is unclear whether the Boland data can truly be taken as evidence that information about instruments is not made available by the verb during sentence processing.

3 Experimental work

In order to determine whether the role of verb-based information during sentence processing involves access to the full range of information associated with the verb, we need to examine the case for activation for arguments that occur non-adjacent to the verb. At the same time, it will be important to avoid situations that confound the presence of activation for immediately upcoming arguments with the absence of activation for downstream constituents. In the experiments presented here, we use instrument verbs as a way of examining the full extent of verb-based information. As discussed above, Ferretti, McRae and Hatherell (2001) demonstrated that like agents, patients and themes, instruments are closely associated with a verb's lexical representation. However, in spite of this close association between instrument and verb, instruments appear syntactically distant from the verb, and rarely (if ever) occur in a positions adjacent to the verb. Instead, instruments will typically appear separated from the verb by its direct object (example 4a). Furthermore, though instruments are often crucial to the action of a verb, they need not be overtly mentioned in the sentence (example 4b).

2a. Pacey cut the paper with scissors.
2b. Pacey cut the paper.

In this way, instrument verbs are a viable way to test the extent and nature of verb-based activation during sentence processing. Since there is no cause for information about instruments to be linked to processing strategies aimed at anticipating the next upcoming constituent, it becomes reasonable to conclude that if increased activation for instruments occurs at the verb, this constitutes evidence that the verb’s full range of conceptual information is being accessed.

3.1 Experiment 1

Experiment 1 investigated the activation for instruments during sentence processing. Materials were constructed using pairs of verbs that denoted similar actions, but varied as to whether the action preferentially involved an instrument. For example, the verb “poke” was contrasted with the verb “touch.” In this case, both verbs denote some sort of physical contact. For both verbs, this contact may be enacted using only one’s bare hands, or alternatively, via the use of some intermediary instrument. However, the verb “poke” is more likely to be interpreted as involving the use of the intermediary instrument (like a stick or a pencil), while the verb “touch” is more associated with the use of one’s bare hands. Thus, we predict that if information about instruments is being brought to bear during sentence processing, verbs like “poke” will elicit more looks to potential instruments than will verbs like "touch." Additionally, since the action of "poke" may be performed even without the aid of an instrument, we can insure that looks to the instrument reflect the verb's individual preferences, and are not merely a pragmatic artifact of participants being required to perform an action that by definition, can only be performed by use of an instrument.

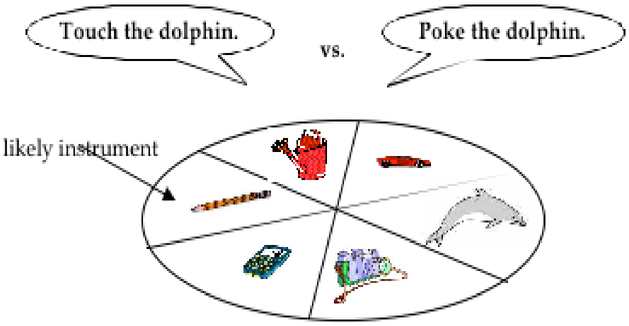


Figure 1: Experiment 1 display and materials

Participants were seated before a display containing six real world objects. For each trial, they would hear a pre-recorded sentence instructing them to manipulate one of the objects in the scene. For critical trials, this instruction contained a member of the instrument/non-instrument verb pair. Participant eye-movements were monitored throughout the trial.

Results showed that in cases where the instruction involved an instrument verb, participants were marginally more likely to look at potential instruments in the display as soon as the verb was encountered ($F_1(1,15)=3.20$, MSE=.01, p=.084, $F_2(1,7)=8.46$, $p<.05$, MSE=.02) and significantly more likely to look at an instrument during the pronuciation of the patient noun in the instruction ($F_1(1,15)=28.33$, $p<.01$, MSE=.13, $F_2(1,7)=31.89$, $p<.01$, MSE=.06). This result obtained even in cases when the participant ultimately chose to perform the action of the instrument verb without the use of an instrument ($t_1(10)=1.99$, $p<.05$). MSE=.03, $t_2(5)=1.68$, p=.08, MSE=.01).

3.2 Experiment 2

Experiment 2 was designed to address councerns that the results of experiment 1 were an artifact of the real-world pragmatic demand of the experimental task rather than the specific information provided by the verbs of the experiment. In particular, there was some legitimate concern that since participants were required to carry out the action, looks to the instrument may have been linked more to the demands of action planning in a real-world context than stemming from the activation of the verb's lexical entry.

Experiment 2 presented the same instrument and non-instrument verb pairs as in experiment 1, but in a context where the participant was not required to perform an action. In this experiment, the participant was seated before a computer screen depicting a person seated at a table containing three objects, including both the item that served as the direct object of the verb, as well as a likely instrument for the action. As they viewed this scene, they heard a sentence describing what was about to occur. The sentences contained the same instrument and non-instrument verbs as tested in experiment 1. Eye movements to the display during the pronunciation of the spoken materials were recorded.

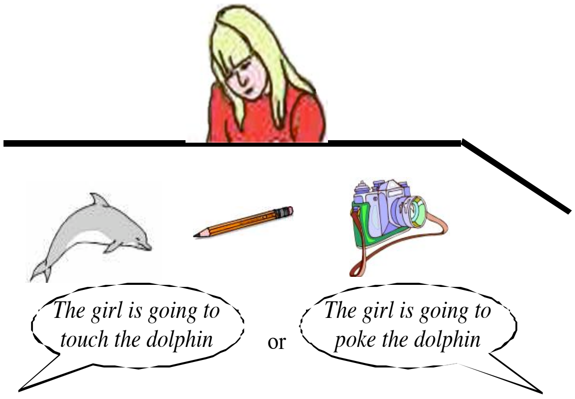

The girl is going to touch the dolphin
or
The girl is going to poke the dolphin

Figure 2: Experiment 2 display and materials

Analyses of eye movements during the windows of time corresponding to the pronunciation of the preverb content, the verb, and the direct object noun phrase revealed that the looks to the display during the verb region varied across the two conditions. Unlike the previous experiment, however, the difference manifested not as a greater amount of looks to the instrument in the display during instrument verb conditions, but rather as a greater number of looks to the direct object item during non-instrument verbs ($F_1(1,68)=5.63$, MSE=.02, $p<.05$, $F_2(1,28)=4.05$, MSE=.01, p=.05). This result is in line with the findings of Altmann and Kamide (1999), which demonstrated that access to the verb engendered immediate looks to likely upcoming patients of the verb in the display. However, the fact that the anticipatory looks to upcoming patients obtained only during non-instrument verb trials raises questions. The increase in looks to the direct object item may reflect an unforseen difference in the predictability of the upcoming direct object across the two conditions, with the object of non-instrument bias verbs in this context being much more predictable than objects of instrument bias verbs in the same context, or alternatively, it may be that the difference between the verbs is directly tied to their instrument status; here, we are theorizing that access to an instrument verb could provide access to information both about its upcoming patient as well as instruments that are likely to participate in the action. Since both of these items are depicted on the screen, we might expect anticipatory looks to be split between the two possible participants. For non-instrument bias verbs, on the other hand, we expect that access to the verb makes available only information about its upcoming patient. In this case, we would expect all anticipatory eye movements generated by access to the verb to focus on the patient item in the display. Thus, it is critical to determine the relative predictability of the upcoming patient item for each verb condition

We conducted a post-hoc survey testing the predictability of the upcoming patient for both instrument and non-instrument bias verbs. Participants saw a printout of the scenes from experiment two, as well as a partial sentence based on the spoken material corresponding to that scene. For the example shown above, the participant would have seen a printout version of the girl and the table full of items, and seen at the bottom a sentence like, "The girl is going to poke the" or "The girl is going to touch the" (depending on condition). Participants were told to circle the item that they felt was most likely to complete the sentence.

For both the instrument and non-instrument verbs, the upcoming patient was highly predictable, with participants choosing this item as the likely sentence continuation on 87% of trials which was significantly more than the instrument and the distractor item ($F_1(1,46)=543.85$, $MSE=.04$, $p<.001$, $F_2(1,14)=274.97$, $MSE=.02$, $p<.001$). Additionally, the percentage of trials where participants chose the patient item as the sentence continuation did not differ across verb type; there was no hint of a main effect of verb type, and no interaction between verb type and display item chosen (Fs<1) These results strongly suggest that the combination of spoken materials and visual displays in experiment 2 made the patient of the verbs' action highly predictable, and furthermore, the degree of patient predictability was the same for both instrument and non-instrument verbs.

In light of the results of experiment 2b, the results of experiment 2 may now be reevaluated. In particular, the high level of patient predictability would lead us to expect a large number of looks to the patient item in the display (following Altmann and Kamide, 1999). And in fact, this is precisely the result we see for the non-instrument verb condition. For instrument verbs, however, we see something very different. Instead of an increase in looks to the upcoming patient, looks to both the patient and the instrument remain relatively equal, receiving 26 and 28 percent of looks, respectively (F's < 1). This result, in turn follows if we assume that access to the instrument-biased verb had made available information about all of the verb's upcoming participants, and thus anticipatory looks are divided between the instrument and the direct object in the display.

3.3 Experiment 3

The results of experiments 1 and 2 both show evidence for verb-based activation of instrument roles, manifested in different ways. While experiment 1 required participants to act upon real-world objects, experiment 2 allowed them to listen to materials while watching a computer screen. It is likely that the difference in evidence for instrument activation stems from this difference in task type. However, experiment 1 also involved six available display items while experiment 1 only involved two. We know that the predictability of the upcoming direct object item plays a large role in the distribution of anticipatory eye-movements. The reduction in set items from six to three may have increased the predictability of the direct object items in experiment 2, in turn accounting for the differences between the results of the two experiments.

Experiment 3 was designed as a direct real-world parallel to experiment 2. Like experiment 1, it required participants to manipulate items in the real world according to spoken instructions while their eye-movements were being tracked. However, unlike experiment 1, only three items were available within the display. These items were the same three pictured in the onscreen displays of experiment 2. Thus, experiment 3 matched experiment 1 in procedure and experiment 2 in direct object predictability.

Early results of experiement 3 match those of experiment 1, with more looks to the instrument during the verb region of instructions containing instrument-biased verbs than during the same region of trials containing non-instrument biased verbs ($F_1(1,60)=4.88$, $MSE=.07$, $p<.05$, $F_2(1,28)=3.19$, $MSE=.03$, $p=.08$). In light of this finding, we can conclude that changes in task type rather than patient predictability account for the different manifestations of instrument activation between experiments 1 and 2.

4 Conclusion

This series of experiments demonstrated that instrument information is immediately available at the verb. This finding provides strong support for models in which the full contingent of a verb's participant information is immediately accesses as soon as the verb is encountered. Additionally, the finding bears upon the standing of instrument roles within a verb's lexical representation. Though instruments may not bear the same type of verb-argument relationships as do more canonical arguments such as themes or patients, their activation upon access to the verb would seem to indicate that instrument information is stored as part of a verb's core representation. This view is in line with models such as the one presented in Koenig, Mauner, and Bienvenue (2003), where lexical encoding of event participants is dependent on a number of separate criteria. This in turn allows for a diverse types and strengths of argument relationships within a verb's lexical entry.

Finally, the results underline the importance of the nature of the experimental task on the form of the outcome. In cases where the participant was required to carry out the action described, activation for instruments was seen in the form of more looks to the instrument in the display. In cases where direct action was not required of the participant, however, activation for instruments was measurable as a suppression of looks to a highly predictable patient item.

5 Acknowledgements

Thanks to our stellar R.A.s Dana Subik, Reena Kanodia, Karen H. Bobrow, and Stephanie Roberts.

6 References

G. Altmann and Y. Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* , 73(3):247-264

J. Boland. 2002. Visual Arguments. Submitted.

J. Boland, M. Tanenhaus, S. Garnsey, and G. Carlson. 1995. Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language*, 34:774-806.

B. Dosher and A. Corbett 1982. Instrument inference and verb schemata. *Journal of Verbal Learning and Verbal Behaviour*. 10:531-539

T. Ferretti, K. McRae, and A. Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language,* 44:516-547

S. Garnsey, N. Pearlmutter, E. Myers, and M. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of*

S. Garrod and A. Sanford. 1981. *Bridging inferences and the extended domain of reference*. In "Universals in linguistic theory" J. Long and A. Baddely ed., pages 1-90, Holt, Rinehart, and Winston, New York.

M. Hare, K. McRae, and J. Elman. 2003. Sense and Structure: Meaning as a determinant of verb subcategorization preferences. Journal of Memory and Language. 48(2):281-303.*Memory and Language*, 37:58-93

S. Kear, and G. Wilson (2000). Familiarity breeds contempt (and it also explains long reading times for references to strongly implied instruments). Paper presented at AMLaP 2000 in Leiden, the Netherlands.

J.-P. Koenig, G. Mauner, and B. Bienvenue 2003. Arguments for Adjuncts. *Cognition*, 89:67-103.

M . Lucas, M. Tanenhaus, and G. Carlson, 1990. Levels of representation in the interpretation of anaphoric reference and instrument reference. *Memory and Cognition,* 18:611-631

J. Saffran, R. Aslin, and E. Newport. 1996. Statistical learning by 8-month old infants. *Science*. 274:1926-1928.

M. Singer. 1979. Process of inference in sentence encoding. *Memory and Cognition*. 7:1191-1210.

J. Trueswell, M. Tanenhaus, and C. Kello. 1993. Verb-specific constraints in sentence processing: Separating the effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning Memory and Cognition*. 19(3):528-553.

Towards Automatic Verb Acquisition from VerbNet for Spoken Dialog Processing

Mary Swift
Department of Computer Science
University of Rochester
Rochester, NY 14607 USA
swift@cs.rochester.edu

Abstract

This paper presents experiments on using VerbNet as a resource for understanding unknown verbs encountered by a spoken dialog system. Coverage of unknown verbs in a corpus of spoken dialogs about computer purchasing is assessed, and two methods for automatically integrating representations of verbs found in VerbNet are explored. The first identifies VerbNet classes containing verbs already defined in the system, and generates representations for unknown verbs in those classes, modelled after the existing system representation. The second method generates representations based on VerbNet alone. The second method performs better, but gaps in coverage and differences between the two verb representation systems limit the success of automatic acquisition.

1 Introduction

TRIPS (The Rochester Interactive Planning System) is a collaborative dialog assistant that performs full loop intelligent dialog processing, from speech understanding and semantic parsing through intention recognition, task planning and natural language generation. In recent years the system has undergone rapid expansion to several new domains. Traditionally the system has used a hand-constructed lexicon, but increased demand for coverage of new domains in a short time period together with the availability of online lexical resources has prompted investigation into incorporating existing lexical resources.

The ability to handle spontaneous speech demands broad coverage and flexibility. Verbs are a locus of information for overall sentence structure and selectional restrictions on arguments, so their representation and organization is crucial for natural language processing.

There are numerous approaches to verb classification. For example, Levin (1993) defines semantic verb classes that pattern according to syntactic alternations. The Levin classes are the basis of the online lexical resource VerbNet (Kipper, Dang and Palmer, 2000; Kipper 2003). However FrameNet (Baker, Fillmore and Lowe, 1998), another hand-crafted lexical resource, classifies verbs using core semantic concepts, rather than syntactic alternations (see Baker and Ruppenhofer (2002) for an interesting comparison of the two approaches). Machine learning techniques have been used to induce classes from distributional features extracted from annotated corpora (e.g., Merlo and Stevenson, 2001; Schulte im Walde, 2000).

This paper reports experiments on using VerbNet as a resource for verbs not defined in TRIPS. VerbNet coverage of unknown verbs occurring in a corpus of spoken dialogs about computer purchasing is evaluated. VerbNet coverage has been previously evaluated in (Kipper et al, 2004b) by matching syntactic coverage for selected verbs in PropBank (Kingsbury and Palmer, 2002). In the present evaluation, TRIPS obtains representations from VerbNet for use during parsing to automatically generate semantic representations of utterances that can be used by the system to reason about the computer purchasing task.

The experiments explore methods for automatically acquiring VerbNet representations in TRIPS. The verb representations in TRIPS and VerbNet were developed independently and for different purposes, so successfully integrating the two presents some challenges. Verb classification in TRIPS is organized along semantic lines similar to FrameNet (see section 2) instead of the diathesis-based classification of VerbNet. Dzikovska (2004) has noted that there is a good deal of overlap between the two in terms of the representation of predicate argument structure and

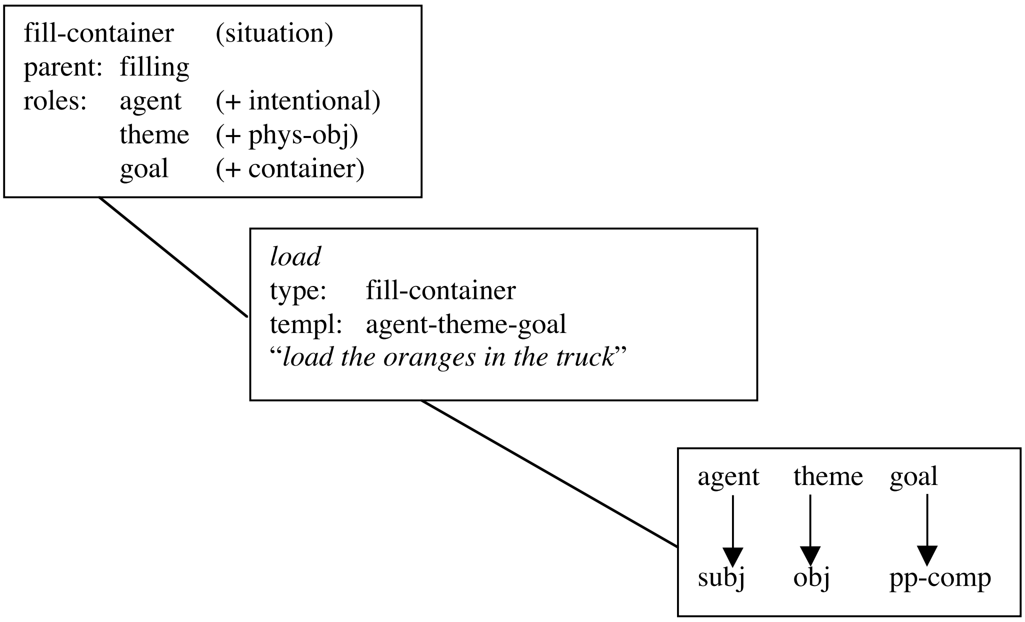


Figure 1: Schematic of the three main components of a TRIPS lexical definition: semantic type, lexical entry, and linking template for one sense of the verb *load*.

associated thematic roles. The experiments reported here provide a more detailed comparison between the two systems and show that in spite of the similarities, there are enough differences to make the integration challenging.

Two automatic acquisition methods are explored. The first creates definitions for verbs in VerbNet classes containing verbs already defined in TRIPS, using the existing definition as a model. The second method generates lexical definitions based on VerbNet information alone. The methods are evaluated by integrating the new definitions into the system and parsing a corpus of transcribed utterances containing the new verbs. Deriving verb definitions directly from VerbNet provides a greater number of acceptable definitions than basing new definitions on existing representations in TRIPS, highlighting some of the difficulties in reconciling independently developed verb representation systems.

2 Verb representation in TRIPS

A lexical representation in TRIPS consists of an orthographic form, part of speech, morphological specifications (if the standard paradigm does not apply), and sense definitions. A lexeme may have one or more sense definitions, which consist of a semantic type with associated thematic roles and semantic features (Dzikovska 2004), and a template that specifies the linking between the thematic roles and syntactic arguments.

The current semantic verb hierarchy takes FrameNet frames (Baker, Fillmore and Lowe, 1998) as its starting point, but incorporates characteristics that streamline it for use in practical spoken dialog processing, such as hierarchical structure and a reduced set of role names (Dzikovska, Swift and Allen, 2004). Each sense definition also includes an example of usage and a meta-data vector that records the origin and date of entry, date of change, and comments. A (simplified) schematic representation for the definition for the verb *load* is shown in Figure 1.

The semantic hierarchy classifies verbs in terms of semantic types that describe the predicate-argument structure. Syntactic frames for licensed constructions are not part of the class specification, as they are in VerbNet. Rather, they are enumerated in the lexical entry itself, as a component of a sense entry.

At the time of this evaluation there are 522 verb lemmas in the TRIPS lexicon. Roughly half of these are also found in VerbNet, although the sense distribution for identical lemmas do not always correspond, as the evaluation in section 4 shows.

3 VerbNet

VerbNet is a hierarchical verb lexicon that uses the Levin verb classes to systematically group verbs into "semantically coherent" classes according to the alternations between the different syntactic frames in which they appear. VerbNet expands on the Levin classes by providing explicit syntactic and semantic information, including thematic roles annotated with semantic features, and syntactic frames for each verb. VerbNet frames use the thematic role names to describe the syntactic constructions in which a verb can appear. For example, the frame Agent V Patient describes a transitive construction for change of state verbs, as in *Floyd broke the vase*.

The experiments reported here are based on VerbNet v1.5,[1] consisting of 191 main classes subsuming more than 4000 verb senses and approximately 3000 verb lemmas.

4 Evaluation

A new corpus in a computer purchasing domain is used for the evaluation. The corpus data consist of human-human dialogs collected as a basis for development of a new computer purchasing domain. The interlocutors model a scenario in which users interact with an intelligent assistant to purchase computer equipment. The corpus comprises 23 dialogs totalling approximately 6900 utterances. At the time of the evaluation there are 139 verbs in the computer purchasing corpus that are not defined in TRIPS (henceforth 'target verbs'). Of these, 66 have definitions in VerbNet.

Two methods (described in sections 4.1.1 and 4.1.2) were used to automatically acquire target verb definitions from VerbNet, which were then used to parse a test corpus of transcribed utterances in which the target verbs occur extracted from the computer purchasing corpus.

4.1 Method

The primary test set focuses on the 49 target verbs in VerbNet that are in classes that also contain TRIPS exemplars: *accelerate*, *admit*, *bet*, *bump*, *clog*, *concern*, *count*, *detect*, *differ*, *disappoint*, *expand*, *filter*, *fold*, *freeze*, *grow*, *guarantee*, *install*, *intend*, *invest*, *investigate*, *knock*, *lean*, *listen*, *melt*, *oppose*, *overwhelm*, *paste*, *plug*, *print*, *punch*, *render*, *roll*, *sacrifice*, *satisfy*, *scan*, *serve*, *settle*, *shop*, *spill*, *stick*, *strip*, *subtract*, *suffer*, *surprise*, *tack*, *tempt*, *void*, *weigh*, *wrap*.

A test corpus of 82 transcribed utterances containing instances of target verbs was extracted from the main corpus. In some cases there is a single instance of a target verb, such as *void* in *That voids the warranty*, while other verbs appear frequently, as is the case with *print*.

For the evaluation, the test corpus was parsed with two different versions of the lexicon, one that included target verb definitions based on existing TRIPS structures and the other included target verb definitions based on VerbNet data alone.

When target verb representations were not based on a TRIPS class match, representations for 17 additional verbs were generated: *advance*, *exit*, *fax*, *interest*, *package*, *page*, *price*, *rate*, *set*, *slow*, *split*, *supply*, *support*, *train*, *transfer*, *wire*, *zip*. These verb representations were evaluated on a separate corpus of 32 transcribed target utterances.

4.1.1 Acquiring verbs based on TRIPS representations

The first method automatically generated verb definitions for the target words by identifying VerbNet classes that contained verbs for which definitions already existed. If a VerbNet class contained a verb already defined in TRIPS, the frames associated with the VerbNet class were compared to the linking templates for all senses defined for the TRIPS verb. If a match was found, lexical entries based on the existing representations were generated for the target verb(s) in that VerbNet class. The new verbs were defined using existing semantic types, their associated thematic roles, and the linking template(s) corresponding to the matching sense entry. An example of a successful match is target verb *subtract* found in VerbNet class remove-10.1, which includes the frames Agent V Theme and Agent V Theme (prep src)[2] Source. The verb *remove* is in this class, and it is also defined in TRIPS as semantic type REMOVE with the roles Agent, Theme and Source.

Although 49 target verbs are in VerbNet classes that contained TRIPS exemplars, this method resulted in just 33 target verb definitions since the frame comparison procedure failed to find a sense match for several of the target verbs.

Identifying a sense match for a given verb by matching linking templates to VerbNet syntactic frames is not straightforward (see also Kipper et al. (2004a, 2004b) for a similar discussion of issues in matching VerbNet and PropBank representations). The verb classes and associated roles used in the two systems were developed independently and for different purposes. Currently TRIPS distinguishes 30 roles for verbs,[3] and VerbNet distinguishes 21 (Kipper 2003). TRIPS roles and their (potentially) corresponding VerbNet roles are listed below.

[1] www.cis.upenn.edu/group/verbnet/download.html

[2] A class of prepositions that can introduce a Source.

[3] Only roles that appear in the linking templates for verbs are discussed. TRIPS also assigns role names to common general modifying phrases (for example, the *for* phrase in *He studied for the test* is assigned the role Reason) and distinguishes roles for nouns, adverbs, and adjectives to aid in parsing and interpretation (see Dzikovska (2004) for discussion).

TRIPS	VerbNet
Core	
Addressee	Recipient
Agent	Agent, Actor(1)
Beneficiary	Beneficiary
Cause	Agent
Cognizer	Agent, Experiencer
Experiencer	Experiencer
Instrument	Instrument
Recipient	Recipient
Theme	Theme(1), Patient(1), Cause, Stimulus
Spatial Location	
Container	Location
Goal	Destination/Location
To-loc	Destination/Location
Source	Source/Location
From-loc	Source/Location
Path	Location
Spatial-loc	Location
Trajectory	
Along	--
Via	--
Co-Roles	
Co-Agent	Actor2
Co-Theme	Theme2, Patient2
Sentential complements (primarily)	
Action	--
Effect	--
Other	
Affected	Patient
Assoc-Info	Topic
Cost	Asset, Extent
Part	--
Predicate	Predicate
Property	Attribute
Result	Product
Time-Duration	Time

The mid-level thematic roles (cf. semantic roles that are frame-specific, such as those used in FrameNet, and macrorole cluster concepts such as Dowty's (1990) Proto-Agent and Proto-Patient) used in TRIPS and VerbNet are difficult to apply consistently, especially on a large scale.[4] Attempts to use one such system to predict another can be problematic. In many cases TRIPS and VerbNet role correspondences are not unique. For example, TRIPS distinguishes a Cognizer role but VerbNet does not – for the verbs *think*, *believe*, and *assume*, the TRIPS Cognizer role corresponds to the VerbNet Agent role, but for the verb *worry*, the TRIPS Cognizer role corresponds to the VerbNet Experiencer role. Conversely, VerbNet makes role distinctions that TRIPS does not, such as Theme and Patient. Furthermore, in the case of identical role names, parallel usage is not assured. For example, TRIPS and VerbNet both distinguish a Cause role but use it in different ways. In TRIPS the Cause role is used as a non-intentional instigator of an action, i.e. "Causer", while in VerbNet it is used as the "Causee", e.g., as the role of *the thunderstorm* in *Spike fears thunderstorms*. In another case, the TRIPS Instrument role is required to be a physical object, while VerbNet has a broader usage, as it assigns the Instrument role to *A murder* in *A murder began the book*.

Another difference of the TRIPS role system is the assignment of thematic roles to certain phrases in a verb's subcategorization frame that have no corresponding role in traditional thematic role schemes. For example, TRIPS identifies sentential complements with role names such as Action for the verbs *try* and *want*. In addition, TRIPS has a more finely articulated role set than VerbNet for locations and paths. TRIPS distinguishes roles such as Along for the trajectory of an action, as in *Route 31* in *The truck followed Route 31 to Avon* and Via for the location through which a motion trajectory (potentially) passes, as in *Avoid the mountains*.

Additional complexities are introduced into the frame matching task for prepositional complements (see Kipper et al., 2004a).

4.1.2 Acquiring verbs based on VerbNet representations

The second method for generating new target verb definitions used VerbNet data alone to generate the semantic type, thematic roles and linking templates necessary for the TRIPS lexical representation. For every VerbNet class containing a target verb, a new semantic type was defined using the VerbNet class name and roles as the type label and the associated thematic roles. The linking templates were generated from the VerbNet frames, which include syntactic category and thematic role information.

This method generated definitions for all 49 of the target verbs found in VerbNet, as well as for the additional 17 target verbs that appear in VerbNet, but in classes that did not include verbs defined in TRIPS.

4.2 Results

The two methods for generating new verb entries were evaluated by integrating the target verb definitions into the system (independently, in two conditions) and then parsing test utterances derived from the computer purchasing domain. The

[4] PropBank (Kingsbury and Palmer, 2002) eschews such thematic role labels altogether, using numbered place holders such as Arg0 and Arg1.

analyses generated by the parser were then scored for accuracy. For the parser representation of an utterance to be counted as accurate, the analysis must contain both an appropriate sense (semantic type) for the target verb and correct role assignments for its arguments. The results are shown in Table 1.

A greater number of acceptable verb representations were obtained by generating entries directly from VerbNet rather than trying to base them on an attempted match with existing TRIPS structures. This is in part due to the complexity of the matching process, and also because of the relatively small number of verbs in TRIPS. Only a few target verbs were successfully matched with the first method, such as *expand* in *You might want to expand it*, which was classified with TRIPS semantic type ADJUST, and has the roles of Agent and Theme.

Data	# verbs	Method	Utts	Acc
Target verbs with TRIPS exemplars	49	I: TRIPS	82	11%
Target verbs with TRIPS exemplars	49	II: VN	82	37%
Extra target verbs from VN	17	II: VN	32	37%

Table 1: Results for parsing test utterances with new verb definitions

The results indicate that it is somewhat easier to generate new linking templates based on VerbNet information than trying to match them with existing structures in TRIPS. Using VerbNet data alone, successful interpretations for a number of prepositional complements are generated, such as *What* (Oblique) *are you* (Experiencer*) interested* (amuse-3.1*) in*? However, in the interpretation of *He spilled coffee on it*, *coffee* is assigned to a location role. This type of error could be corrected by incorporating semantic features for selectional restrictions on argument selection, which are included in VerbNet, and integrating them into the lexical definitions is planned for future work. However, TRIPS has its own system of semantic features for the same purpose so additional analysis required before the VerbNet feature representation can be fully integrated with TRIPS.

In some cases there were idioms in the data for which a correct analysis could not reasonably be expected. For example, the target verb *roll* was reasonably mapped to the MOVE semantic type with the first method, but the instance of *roll* in the test corpus is an idomatic one, as in *Let's roll with that*, and the system incorrectly assigns *that* to an Instrument role. Predictably, neither method yielded an appropriate sense for this case, nor for other idiomatic usages such as *Let's stick with the twelve-inch powerbook*.

Missing senses and frames in VerbNet were an additional source of error for both methods of verb definition generation. For example, VerbNet lacks a frame with a sentential complement for *tempt*, as in *I'm tempted to get the flat screen monitor*. Another case of missing sense is for the target verb *support*, as in *That capability is only supported in Windows XP*. *Support* is found in two VerbNet classes, contiguous_location-47.8 and admire-31.2, neither of which are appropriate in this case.

The evaluation revealed that several of the target verbs occurred together with particles, such as *punch in* as in *Let me just punch in those numbers*, as well as *bump up*, *clog up*, *fold up*, *knock off, knock down*, *plug in*, *punch in*, *set up*, *split up*, *slow down*, and *wrap up*. These were a major source of error in this evaluation since particle verbs are not generally represented in VerbNet.[5] 16 utterances from the primary target test corpus contain particle verbs, and failure to handle them accounts for 31% of the error for the condition in which the VerbNet derived definitions are tested. 7 utterances in the test corpus for the extra verbs contain particle verbs and these account for 35% of the error for that test set.

5 Summary and Conclusion

It had seemed that using TRIPS representations to model new verbs would yield better results, since in principle more of the information built into TRIPS could be used, but this turned out not to be the case. This method could be improved with additional comparative analysis along with expansion of the TRIPS lexicon, but there will still be enough differences to pose difficulties for automatic mapping between the systems. Automatically generating representations from VerbNet data alone produced better results, but adopting VerbNet classifications wholesale is impractical as they are not always an appropriate level of semantic representation for the parsing and reasoning performed by the system. For example, the class other_cos.45.4 has more than 250 members. Even though they are all change of state verbs, efficient parsing and effective reasoning require finer-grained distinctions to process

[5] The clustering analysis reported in Kingsbury and Kipper (2003) identifies particle verbs, such as *pull out*, compatible with certain VerbNet classes.

meanings as disparate as, for example, *unionize* and *vaporize*.

The ability to use VerbNet representations directly is still only a partial solution to expanding the system's verb coverage. For these experiments, less than half of the unknown verbs were actually found in VerbNet. Verbs not found include *aim*, *apply*, *compromise*, *concentrate*, *customize*, *discuss*, *elaborate*, *format*, *manipulate*, *optimize*, *program*, *scroll*, *subscribe*, and *troubleshoot*. Of the target verbs found in VerbNet, an appropriate sense was not always represented.

The Levin verb classes are not exhaustive and focus on noun phrase arguments and prepositional complements, so for example verbs with sentential complements are underrepresented, although VerbNet has extended and modified the original classes on which it is based, and continues to be refined (Kipper et al., 2004a). There are still systematic gaps, most importantly for this evaluation, particle verbs.

With its rich syntactic and semantic representation, VerbNet promises to be a useful resource for extending lexical coverage in TRIPS. VerbNet representations also include links to corresponding senses in WordNet (Fellbaum 1998), which strengthens the network of lexical information available that can contribute to better handling of unknown words when they are encountered by the system. However, achieving a representation that combines the predictability of syntactic alternations together with the level of semantic classification needed for spoken dialog processing remains a challenge.

6 Acknowledgements

I thank John Dowding, Myroslava Dzikovska, Dan Gildea and two anonymous reviewers for helpful comments and discussion. This material is based upon work supported by a grant from The National Science Foundation award #IIS-0328811, DARPA under grant number NBCHD030010 via subcontract to SRI number 03-000223 and The National Science Foundation under grant number E1A-0080124. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of above named organizations.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING-ACL-1998*, Montreal, CA.

Collin F. Baker and Josef Ruppenhofer. 2002. FrameNet's Frames vs. Levin's Verb Classes. In *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, J. Larson and M. Paster (eds.), pages 27-38, Berkeley, CA.

David Dowty. 1991. Thematic Proto-roles and Argument Selection. *Language* 67(3).

Myroslava Dzikovska. 2004. *A Practical Semantic Representation for Natural Language Parsing*. Ph.D. thesis, University of Rochester.

Myroslava Dzikovska, Mary Swift, and James Allen. 2004. Building a Computational Lexicon and Ontology with FrameNet. In *Workshop on Building Lexical Resources from Semantically Annotated Corpora at LREC-2004*, Lisbon.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communications. MIT Press, Cambridge, Massachusetts.

Paul Kingsbury and Karin Kipper. 2003. Deriving Verb-Meaning Clusters from Syntactic Structure. In *Workshop on Text Meaning at HLT-NAACL*, Edmonton, Canada.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *LREC-2002*, Las Palmas, Spain.

Karen Kipper Schuler. 2003. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis proposal, University of Pennsylvania.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *AAAI-2000*, Austin TX.

Karin Kipper, Benjamin Snyder, and Martha Palmer. 2004a. Using Prepositions to Extend a Verb Lexicon. In *NAACL-2004*, Boston.

Karin Kipper, Benjamin Snyder, and Martha Palmer. 2004b. Extending a verb-lexicon using a semantically annotated corpus. In *Workshop on Building Lexical Resources from Semantically Annotated Corpora at LREC-2004*, Lisbon.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3), September.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behavior. In *COLING-2000*, Saarbrücken, Germany.

The Processing of Regular and Irregular Verbs

Wieke M. Tabak
Radboud University of Nijmegen
Wundtlaan 1
6525 XD Nijmegen,
The Netherlands,
wieke.tabak@mpi.nl

Robert Schreuder
Max Planck Institute
Wundtlaan 1
6525 XD Nijmegen,
The Netherlands,
rob.schreuder@mpi.nl

R. Harald Baayen
Radboud University of Nijmegen
Wundtlaan 1
6525 XD Nijmegen,
The Netherlands,
harald.baayen@mpi.nl

Abstract

This paper provides an overview of how a series of different distributional properties of irregular and regular verbs affect lexical processing in single-word comprehension and production. (Tabak et al., 2005) show that it is possible to predict whether a verb is (ir)regular from not only frequency, but also from its neighborhood density, inflectional entropy, morphological family size, number of synsets, its auxiliary, and its number of argument structures. These variables were observed to be predictive for both response latencies and errors in a visual lexical study of 286 Dutch verbs. Interestingly, the greater number of synsets characterizing irregulars led to especially short response latencies for irregular past plurals. Moreover, a higher information complexity, as estimated by the inflectional entropy measure developed in, led to shorter response latencies, and especially so for irregular verbs. In this study, we investigated whether such measures of semantic density could be observed to play a role in word naming also. Two word naming experiments were carried out, simple word naming as well as cross-tense naming. Semantic variables were predictive primarily in cross-tense naming, a task which also revealed effects suggesting competition between the form read and the form said. This competition challenges dual route models of production (Pinker, 1991; Pinker, 1999) and argues for exemplar models of direct lexical access.

1 Introduction

It is widely believed (Pinker, 1999; Pinker and Ullman, 2002) that the difference between regular and irregular verbs is restricted to form. However, recent studies (Ramscar, 2002; Patterson et al., 2001) suggest that semantic and contextual factors are also relevant for understanding how regular and irregular verbs are processed. A lexical statistical survey (Baayen and Moscoso del Prado Martín, 2004) revealed, furthermore, that regulars and irregulars differ in semantic density: Irregulars have denser semantic networks than regulars. For instance, when regulars and irregulars are matched for frequency, irregular verbs tend to have more meanings than regular verbs. Regular and irregular verbs also tend to have different aspectual properties, as witnessed by the non-uniform distribution of auxiliary verbs in Dutch and German. Regulars favor *hebben*, 'have', while irregulars favor *zijn*, 'be', the auxiliary marking telicity. Regular and irregular verbs are also non-uniformly distributed over Levin's verb argument alternation classes (Levin, 1993).

Another dimension on which regular and irregular verbs differ is the information complexity of a verb's paradigm, as estimated by the inflectional entropy measure in Dutch (Moscoso del Prado Martín et al., 2004). Irregular verbs tend to have somewhat reduced inflectional entropies compared to regulars.

(Tabak et al., 2005) probed the consequences of these distributional differences for the comprehension of regular and irregular verbs in Dutch with visual lexical decision. They used a regression design with 143 regular and 143 irregular morphologically simple verbs, and contrasted number (singular versus plural) and tense (present versus past). Examples of the verb forms involved are shown in Table 1.

Tabak et al. observed facilitatory main effects of the frequency of the inflected form (obtained from the CELEX lexical database (Baayen et al., 1995) henceforth surface frequency), the verb's morphological family size (Schreuder and Baayen, 1997) (calculated from the morphological parses in the CELEX database), and the number of synsets (Baayen and Moscoso del Prado Martín, 2004) in which the verb is listed in the verb section of the Dutch EuroWordNet (Vossen et al., 1999), the Dutch version of the English WordNet developed by Miller and colleagues (Miller, 1990). In addition, a facilita-

loop	singular	present tense	irregular
liep	singular	past tense	irregular
loop-en	plural	present tense	irregular
liep-en	plural	past tense	irregular
lach	singular	present tense	regular
lach-te	singular	past tense	regular
lach-en	plural	present tens	regular
lach-ten	plural	past tense	regular

Table 1: Examples of regular and irregular verb forms in Dutch.

tory effect of inflectional entropy was observed. Unsurprisingly, regular verbs elicited shorter response latencies than irregular verbs. More interesting are two interactions with regularity. First, the synset measure revealed a greater facilitatory effect for the past plural forms of irregular verbs. Apparently, these forms, which are both irregular and morphologically complex, benefitted most from the higher semantic density characterizing irregulars. Second, the facilitatory effect of inflectional entropy was stronger for irregulars than for regulars.

(Ullman, 2001) reported frequency effects for regular verbs that rhyme with irregular verbs, and the absence of frequency effects for regular verbs that do not do so. Tabak and colleagues therefore investigated whether there might be a similar dissociation for the regular verbs in their experiment. What they found was that regulars that rhyme with one or more irregulars elicited longer response latencies than non-rhyming regulars. However, non-rhyming regulars exhibited a frequency effect just as the rhyming regulars.

These results show that the comprehension of regular and irregular verbs is more intricate and interesting than suggested by the dual route model of Pinker and colleagues. The frequency effects for regulars point in the direction of exemplar based models (Bybee, 2001), and the interactions of semantic measures with regularity bear witness to the greater semantic entanglement of irregular verbs (Patterson et al., 2001).

2 Word Naming

The next question to be addressed is whether semantic effects can be observed as well in word naming, a task that adds production of the verb form to visual recognition. It is well known that semantic effects are reduced in word naming compared to visual lexical decision (Balota et al., 2003). Hence, it is far from self-evident that the greater effect of semantic density of irregulars would persist in this task. We therefore ran a simple word naming experiment with the same materials as in the lexical decision experiment summarized above, in which participants were asked to say aloud the singular and plural forms in the present and past tense presented on the computer screen. Pronunciation latencies were measured from word onset.

The naming paradigm that has figured most prominently in the debate about regular and irregular verbs requests subjects to name the past tense form after having been presented with the corresponding present tense form. This task asks subjects to proceed from the present to the past, presumably because this would reflect the way in which the marked past tense form is derived or accessed from the unmarked present tense form. In the light of the evidence that has been accumulating in recent years that lexical storage is not restricted to irregular inflected words (Baayen et al., 2002; New et al., 2004), the idea that a regular past tense form would only be accessible through its present tense form becomes unattractive. In a non-derivational, exemplar-based approach to lexical access, the directionality imposed by the cross-tense naming task taps into meta-linguistic skills that do not necessarily reflect the normal processes of lexical access. In fact, reversing the directionality of the task, by asking subjects to name the present tense form when presented with the past tense form, should lead to a very similar pattern of results. In order to test this possibility, we ran two cross-tense naming experiments, one going from the present to the past, and one from the past to the present.

In simple word naming, it is possible to proceed from the orthography to the phonology and from there to pronunciation without paying much attention to word meaning. Cross-tense naming is a more complex task, in which we expected the role of word meaning to be more substantial. In addition, an exemplar-based approach predicts competition in cross-tense naming between the form seen on the computer screen and the form to be produced, irrespective of regularity.

The results of the two naming experiments are summarized in Figure 1, which illustrates the partial effects of the covariates. (A partial effect is the effect of a variable when all the other variables in the model are held constant.) The grey curves represent the effects observed in the simple naming experiment, the black curves vi-

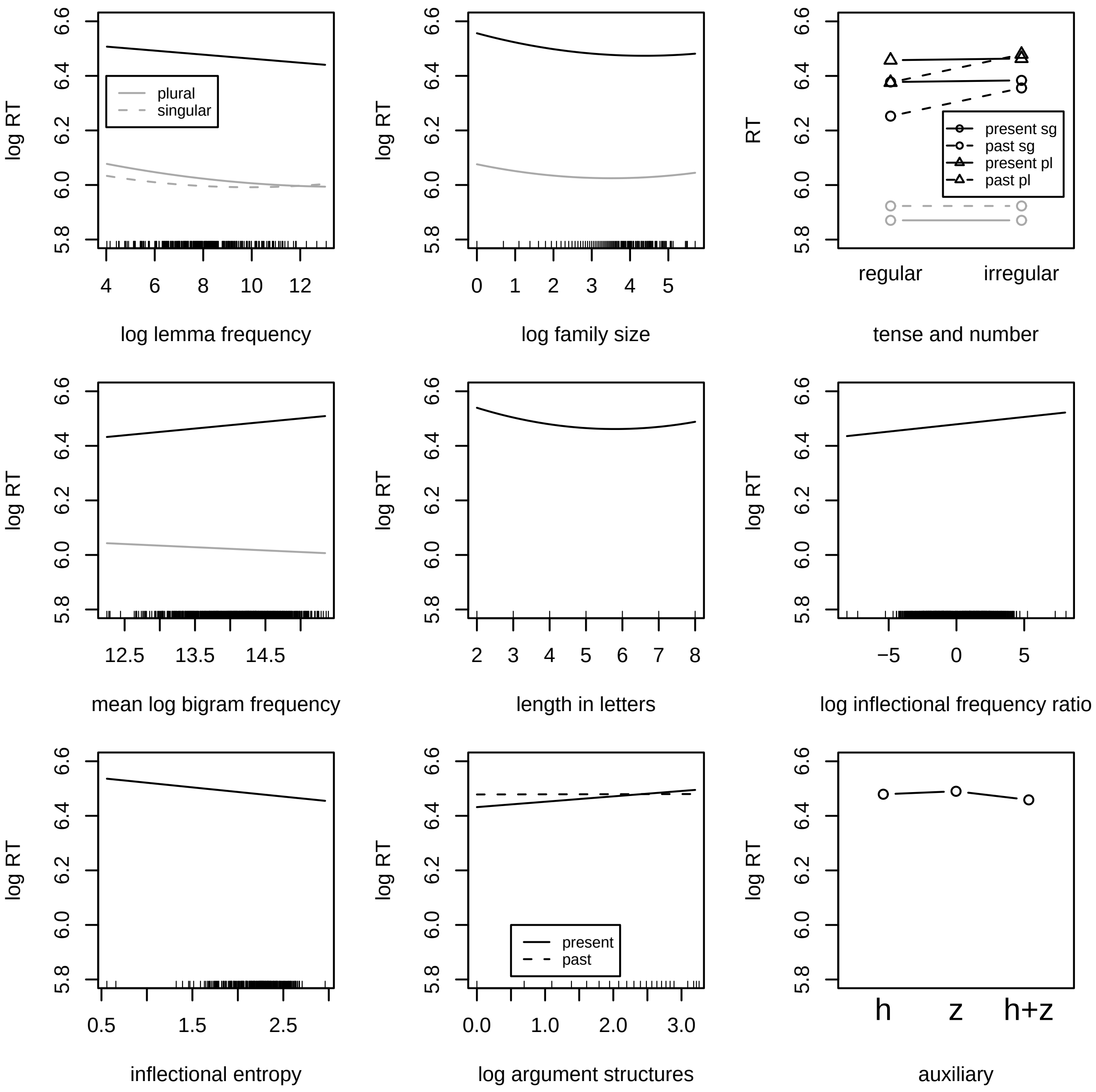


Figure 1: Partial effects of the covariates in the logistic regression model fit to the response latencies of the simple (grey curves) and cross-tense (black curves) naming experiments.

sualize the effects in the cross-tense naming experiment.

A striking difference between the two naming experiments is that cross-tense naming latencies are predicted from a much larger number of variables than is the case for simple word naming. The only predictors (represented by grey lines) for simple word naming are frequency (upper left panel), family size (upper middle panel), tense (upper right panel), and geometric mean bigram frequency (first panel on second row). As expected, higher-frequency words elicited shorter naming latencies (especially for plurals), as did words with bigger morphological families. The latencies for past tense forms were longer than those for present tense forms, irrespective of number or regularity. Finally, a higher bigram frequency corresponded with shorter reaction times. In short, simple word naming did not provide any evidence for a role of semantic measures interacting with regularity. It is noteworthy that there is no interaction of regularity by frequency. This indicates that the frequency effect works the same for both regular and irregular verbs.

The results obtained in cross-tense naming (the black lines in Figure 1) were much richer. We note, first of all, that the response latencies are much longer than in simple naming, which reflects the greater difficulty of the task. The first two panels on the upper row show that we have both a facilitatory frequency effect (for regulars and irregulars) as well as a facilitatory family size effect.

The upper right panel plots the effect of number and tense for regular and irregular verbs. The lines are labeled with the form that subjects had to say, not the forms they saw on the screen. Regular words elicited shorter responses when subjects were pronouncing the paste tense form (on presentation of the present tense form), compared to when they were pronouncing the present tense form (on presentation of the past tense form). No such difference emerged for the irregular verbs. In other words, when subjects proceed from the regular unmarked form to the regular marked form, they are faster than when they have to proceed in the reverse direction. These faster naming times may be the consequence of metalinguistic skills acquired at school, where inflectional paradigms are typically taught going from the present to the past. The faster naming latencies may also simply reflect markedness relations, indicating stronger links from the unmarked to the marked forms than from the marked to the unmarked forms. The shorter naming latencies observed for regular past tense forms compared to their irregular counterparts challenges the dual route model, according to which irregulars would only require lexical look-up in associative memory, while regulars would require both look-up, and upon failure of look-up (the regular inflected form would not be stored), rule-driven affixation. Finally, the upper right panel shows that singulars (represented by circles) were named faster than the plurals (represented by triangles), as expected given the greater morphological complexity of the plural form, which always contains a plural suffix, irrespective of regularity.

A higher geometric mean bigram frequency (calculated for the wordform to be articulated) led to increased naming latencies, as shown in the first panel on the second row. This inhibition is independent of the direction of the inflectional operation required, and is suggestive of some kind of form competition, possibly between the bigram transitions of the form seen (or whatever their mental correlates might be) and the transitions of the form to be produced. If our interpretation is correct, this inhibitory effect supports the hypothesis that the cross-tense naming task involves competition between inflectional variants. Note that in simple naming, bigram frequency is facilitatory instead of inhibitory, as expected for a task in which there is no competition between two word forms.

The central panel shows how the response latencies vary with the length of the presented item. As in visual lexical decision, we observe a u-shaped curve, indicating facilitatory effects for the medium length verb forms.

Recall that we observed an effect of frequency in cross-tense naming. The frequency count that we used here is the (log) lemma frequency available in the CELEX lexical database. This frequency count represents the summed frequencies of all the inflectional variants of a given word. Two other frequency measures that are of potential interest in cross-tense naming are the frequencies of the inflectional form presented on the screen, and the frequency of the inflectional form that has to be pronounced. Obviously, all three frequency measures are highly correlated, and therefore it makes no sense to include all three in a multiple regression analysis. In order to come to grips with the role of the inflectional frequencies, we calculated the ratio of the logged frequency of the form seen to the logged frequency of the form pronounced. We refer to this ratio as the inflectional frequency ratio. It is only mildly correlated with the lemma frequency ($r = 0.176$). We interpret the facilitatory effect of the lemma frequency (shown in the upper left panel of Figure 1) as tapping into the speaker's overall familiarity with especially the meaning of a given verb, following (Bates et al., 2003; Moscoso del Prado Martín et al., 2004; Baayen, 2005). We included the inflectional frequency ratio as a predictor in order to capture potential competition between the form (or inflectional meaning) seen and the form (or inflectional meaning) to be said. We predicted that a greater ratio, as an indication of increased competition between the two inflected forms (or meanings), would be positively correlated with the cross-tense naming latencies.

The last panel on the second row of Figure 1 shows that cross-tense naming latencies indeed increase with this ratio, providing further evidence for lexical competition between the present and past tense forms (or meanings) in cross-tense naming. As there were no interactions with regularity nor with the direction of naming, this finding supports exemplar based models of lexical access. We think, however, that this competition is an artefact of cross-tense naming, and that it is absent in normal production when lexical access is not forced to go to a given inflection via another contextually inappropriate inflection.

The panel in the lower left of Figure 1 plots the facilitatory effect of inflectional entropy. Apparently, subjects benefit from a high inflectional entropy, just as in visual lexical decision.

Another variable that has been observed to be predictive for regularity is the number of argument structures that a verb allows (Baayen and Moscoso del Prado Martín, 2004). For this study, we operationalized this variable by means of a count of the number of argument structures and complementation patterns in a data resource compiled at the Max Planck Institute for Psycholinguistics, Nijmegen. The second panel on the third row of Figure 1 shows that the logarithmically transformed number of argument structures is positively correlated with the response latencies when verbs have to be named in the present tense. This suggests that a verb's argument structures are more likely to be co-activated by the unmarked inflectional tense form.

Our final piece of evidence that semantic structure is mediating production in the cross-tense naming task concerns the auxiliary of the verb. Auxiliary selection in Dutch is determined by the telicity of the event expressed by the verb (Lieber and Baayen, 1997) and (Randall et al., 2003), with telic verbs and verbs construed in a telic event requiring *zijn* instead of the more common *hebben.* The lower right panel plots the effects of the Dutch auxiliaries. Verbs that take both *hebben* or *zijn* were named significantly faster than verbs taking only *zijn* or verbs that take only *hebben*, although the effect was small.

We also investigated the count of verbs that rhyme with a given target verb, following (Ullman, 2001). As this count has very different distributional properties for regular and irregular verbs, we studied its effect for the two kinds of verbs separately. For regular verbs, we distinguished between verbs with and without rhyming irregulars. Unlike (Ullman, 2001), we observed a significant facilitatory effect of lemma frequency for both regular verbs with and without rhyming irregulars. This effect was stronger for regular verbs with rhyming irregulars (which is in line with Ullman's findings for English). At the same time, regular verbs with rhyming irregulars elicited longer naming latencies than regular verbs without rhyming irregulars.

For the set of irregular verbs, the count of rhyming irregulars itself could be included as a measure of the productivity of the vocalic alternation of the verb. As expected, verbs exhibiting more productive vocalic alternation were named faster, irrespective of frequency.

3 Conclusions

The view on the processing of regular and irregular verbs that emerges from our experiments is fundamentally different from the dual route model proposed by (Pinker, 1999) and related studies. This model separates form and meaning, regularity is equated with productivity and rule-based processing, and irregularity with a lack of productivity and storage in memory. Conversely, we have observed the footprints of a much more complex, interconnected, exemplar based system in which inflected forms are accessed directly instead of indirectly through their stems. This system also departs from the standard connectionist view of the production of the past tense as exemplified by the seminal model of (Rumelhart and McClelland, 1986) and subsequent work, which inherited the derivational view of the past tense from generative grammar. The challenge for this new approach to understanding the processing of regular and irregular verbs is to come to grips with the neural architecture and processes that underlie the effects of the measures and variables that we have studied here.

References

R. H. Baayen and F. Moscoso del Prado Martín. 2004. Semantic density and past-tense formation in three germanic languages. *Manuscript submitted for publication, Max Planck Institute for Psycholinguistics.*

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. *The CELEX lexical database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

R. H. Baayen, R. Schreuder, N. H. De Jong, and A. Krott. 2002. Dutch inflection: the rules that prove the exception. In S. Nooteboom, F. Weerman, and F. Wijnen, editors, *Stor-*

age and Computation in the Language Faculty, pages 61–92. Kluwer Academic Publishers, Dordrecht.

R.H. Baayen. 2005. Data mining at the intersection of psychology and linguistics. In A. Cutler, editor, *Twenty-First Century Psycholinguistics: Four Cornerstones*, page Erlbaum, Hillsdale, N.J.

D.A. Balota, M.J. Cortese, S.D. Sergent-Marshall, and D.H. Spieler. 2003. Visual word recognition for single syllable words. *submitted*, X:xx–yy.

Elizabeth Bates, Simona D'Amico, Thomas Jacobsen, Anna Szekely, Elena Andonova, Antonella Devescovi, Dan Herron, Ching Ching Lu, Thomas Pechmann, Csaba Pléh, Nicole Wicha, Kara Federmeier, Irini Gerdjikova, Gabriel Gutiérrez, Daisy Hung, Jeanne Hsu, Gowry Iyer, Katherine Kohnert, Teodora Mehotcheva, Araceli Orozco-Figueroa, Angela Tzeng, and Ovid Tzeng. 2003. Timed picture naming in seven languages. *Psychonomic Bulletin and Review*, 10:344–380.

J. L. Bybee. 2001. *Phonology and language use.* Cambridge University Press, Cambridge.

B. Levin. 1993. *English Verb Classes and Alternations. A preliminary Investigation.* The University of Chicago Press, Chicago.

R. Lieber and R. H. Baayen. 1997. A semantic principle for auxiliary selection in Dutch. *Natural Language and Linguistic Theory*, 15:789–845.

G. A. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–312.

F. Moscoso del Prado Martín, A. Kostić, and R. H. Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94:1–18.

F. Moscoso del Prado Martín, R. Bertram, T. Häikiö, R. Schreuder, and R. H. Baayen. 2004. Morphological family size in a morphologically rich language: The case of finnish compared to dutch and hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1271–1278.

B. New, M. Brysbaert, F. L. Segui, and K. Rastle. 2004. The processing of singular and plural nouns in french and english. *Journal of Memory and Language*, 51:568–585.

K. Patterson, M. A. Lambon Ralph, J. R. Hodges, and J. L. McClelland. 2001. Deficits in irregular past-tense verb morphology associated with degraded semantic knowledge. *Neuropsycologia*, 39:709–724.

S. Pinker and M. Ullman. 2002. The past and future of the past tense. *Trends in the Cognitive Sciences*, 6(11):456–462.

S. Pinker. 1991. Rules of language. *Science*, 153:530–535.

S. Pinker. 1999. *Words and Rules: The Ingredients of Language.* Weidenfeld and Nicolson, London.

M. Ramscar. 2002. The role of meaning in inflection: Why the past tense doesn't require a rule. *Cognitive Psychology*, 45:45–94.

J. Randall, A. van Hout, J. Weissenborn, and R. H. Baayen. 2003. Acquiring unaccusativity: a cross-linguistic look. In A. Alexiadou, E. Anagnostopoulou, and M. Everaert, editors, *The unaccusativity puzzle*, pages 332–353. Oxford University Press, Oxford.

D. E. Rumelhart and J. L. McClelland. 1986. On learning the past tenses of English verbs. In J. L. McClelland and D. E. Rumelhart, editors, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*, pages 216–271. The MIT Press, Cambridge, Mass.

R. Schreuder and R. H. Baayen. 1997. How complex simplex words can be. *Journal of Memory and Language*, 37:118–139.

Wieke Tabak, Robert Schreuder, and R. Harald Baayen. 2005. Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in dutch. In S. Kepser and M. Reis, editors, *Linguistic evidence — Empirical, Theoretical, and Computational Perspectives.* Mouton de Gruyter, Berlin.

M. Ullman. 2001. The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, 30:37–69.

P. Vossen, L. Bloksma, and P. Boersma. 1999. *The Dutch WordNet (CD-ROM).* European Language Resources Association (ELRA), Luxembourg.

Matching Verb Attributes for Cross-Document Event Coreference

Eleftheria Tomadaki and Andrew Salway
Department of Computing, University of Surrey
Guildford, Surrey, UK
GU2 7XH
{e.tomadaki, a.salway}@surrey.ac.uk

Abstract

Collateral texts of different genre can describe the same filmed story, e.g. audio description and plot summaries. We deal with the challenge of cross-document coreference for events by matching verb attributes. Cross document coreference is the task of deciding whether two linguistic descriptions from different sources refer to the same event. This is important for reliable information integration, as well as generating richer machine-executable representations of multimedia material in retrieval and browsing systems. Corpora of audio description and plot summaries were analysed to investigate how they describe the same film events. This analysis shows that events are described by different verbs in the two corpora and has inspired the algorithms for cross-document event coreference, which match verb attributes, rather than verbs themselves. The preliminary evaluation was encouraging, showing a significantly better performance than the baseline algorithm.

1 Introduction

The present era can be characterised by a vast amount of information available in different forms of media; text documents, images, audio and video files etc. Many kinds of electronic information artefacts convey the same story; a fire event, for example, can be broadcast on television or radio, or narrated in a newspaper by the people that were affected; or a fictional story, for example Cinderella can be presented in films, theatre, books, pantomime etc. Information can be conveyed in the form of stories in history, science, current affairs, financial news, fiction etc. The process of narrating a story comprises a sequence of causally connected events organised in space and time. Matching events can be one way to acquire major information about a story.

This research is motivated by the fact that associating information in different texts representing the same story can on the one hand enhance the collection and verification of most available information about one story and more reliable information integration, and on the other hand provide richer machine-executable representations of multimedia material in retrieval and browsing systems, such as film databases.

Natural language textual descriptions can be collateral to a moving image and represent its content in words. Extracting information from collateral text (Srihari, 1995) can address higher levels of semantic video content than video processing alone, as language can express more information than colours, shapes, motion etc. and enhance video indexing, retrieval and browsing. Films entail stories and their content can be described by a range of collateral texts; a story told in a novel can be turned into a film. Novels can total 100,000's words and give detailed descriptions of charaters' cognitive states, which can be expressed by facial expressions in the moving images. Screenplays are the directors' scripts including dialogue, character and setting descriptions as well as instructions to the camera totaling 10,000's words. Audio descriptions are detailed descriptions of the characters' appearance and facial expressions, settings and what is happening on screen at the moment of speaking totaling 1,000's words. Audio description is scripted before it is recorded and includes time-codes to indicate when each utterance is to be spoken, enabling the alignment of the narration with the visual images. Plot summaries narrate the major events of the film in 100's words and include character's desires and goals. The challenge is to understand what is common in different collateral descriptions representing the same events. Consider for example, how the same event (*burned*) for the film English Patient is described in different collateral texts, Figure 1. Each source is heterogeneous, using different vocabulary, grammar structures, amount and kinds of information. These different collateral descriptions can be aligned to audio description fragments, which are temporally associated to the film data;

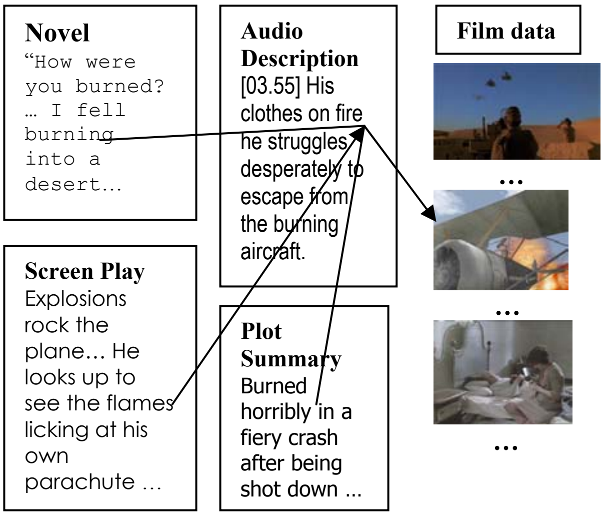

Novel
“How were you burned? … I fell burning into a desert…
Audio Description
[03.55] His clothes on fire he struggles desperately to escape from the burning aircraft.
Film data
…
…
…
Screen Play
Explosions rock the plane... He looks up to see the flames licking at his own parachute ...
Plot Summary
Burned horribly in a fiery crash after being shot down ...

Figure 1: Different collateral descriptions for the same film event.

1.1 Towards Information Integration

A number of terms can describe the process by which information is extracted from different texts relating to the same theme and then associated and combined. The method followed in this work as a first step to integrate event-related information is called *Cross-Document Coreference*; this is the process of deciding whether two linguistic descriptions from different sources refer to the same entity or event and has been applied in specific sets of events, such as election and terrorist events (Bagga and Baldwin, 1999). Recent systems associate entities, extracting nouns and pronouns from different news texts and matching them (Radev and McKowen, 1998). Cross-document coreference appears to be a sub-task of cross-document summarisation by selecting and matching of the crucial information in multiple texts before summarising multiple documents. The task of selecting candidate phrases is expressed in the Document Understanding Conferences (DUC) and is based on the principle of relevance: syntactic patterns are significant, as they describe either a precise and well-defined entity or concise events or situations. Cross Document Structure Theory (CDST) describes several relations included in pairs of matched fragments tested on news articles (Zhang et al, 2003). CDST is tested on relations in homogeneous texts. Related research includes the term information merging describing the process of integrating information about a set of football events, e.g. goal, free kick etc.; the technique applied includes extraction and matching of a set of specific entities, such as football players' names etc from different texts, e.g. tickers, radio transcriptions etc. (Kuper et al, 2003).

Although the two kinds of texts presented in this paper, audio description and plot summaries, describe the same story, they are very different in the vocabulary used, the content and amount of event-related information included; cross-document event coreference in films is perhaps more challenging because it is harder to identify a set of common events.

The goal of the current work is to develop a computational account of how events are expressed in different narrative discourses of the same story in multimedia systems. We focus on the question of how information about an event can be related in different discourses. Our approach is inspired by the corpora analysis, which shows the challenge of matching events in heterogeneous texts, such as plot summary and audio description, as they include different verbs. However, several verb attributes, for instance nouns and proper nouns, are common in both kinds of texts. This analysis has led to the proposal of a method including algorithms that apply event cross-document coreference by matching combinations of verb attributes, rather than matching verbs themselves.

2 Collateral Texts for Films: audio description and plot summaries

Audio description (AD) narrates what is happening on screen for visually impaired people and is available for a range of television programmes, such as series, documentaries, films, children's programmes etc. It is produced by trained experts who follow guidelines while describing, for instance the use of present tense showing that the actions take place at the moment of speaking and the use of proper nouns when there are a lot of participants in a scene to avoid the confusion of the audience. The description is first prepared in electronic format, time-coded and then spoken. The audio description for films is a detailed, long description which involves a story, unfolded in a series of temporally and causally connected events, including characters and plot significant objects, location of the scene, who is speaking, what the characters are doing and wearing, facial expressions and body language, text shown on screen and colours. The following examples are from the audio description for the film English Patient from 3m 40s to 3m 55s:

[03:40] Bullets tear holes in the fuselage.
[03:47] The plane catches fire.
[03:55] His clothes on fire he struggles to escape

In contrast, plot summaries (PS) are short descriptions mentioning the major points of a filmed story, the protagonists and their intentions, locations, time and duration of main events and cause of certain actions. The film is described according to the subjectivity of any author that decides to publish a film summary electronically, without following any guidelines. The following excerpt is from the plot summay for the film English Patient:

> Burned horribly in a fiery crash after being shot down while crossing the Sahara Desert ...

2.1 Corpora Analysis

Two corpora were created to represent and analyse the language used in audio descriptions and plot summaries. The corpora include nine different film categories selected by audio description experts based on the choice of vocabulary, grammar structures and kinds of information conveyed: children's live action and animation, action, comedy, period drama, thriller, dark, romantic and other. The present audio description corpus includes audio description scripts for 56 films, approximately 376,000 words (6,000-8,500 words per script). The current plot summaries corpus includes summaries for the same films (Internet Movie Database), totaling 9,500 words approximately (around 200-400 words per summary). The 100 most frequent words include 41 open class words in the audio description corpus, and 27 open class words in the plot summary corpus. This suggests that audio description and plot summaries are special languages, while comparing them with common language (2 open class words in the first 100 words of the BNC corpus) and other corpora of special languages (e.g. 39 open class words in the linguistics corpus). The most frequent words in both corpora are proper nouns and nouns referring to characters, plot significant objects and time, as well as verbs. However, only a few nouns and proper nouns are the same. In language, an event is typically realised in the form of a verb or noun. We analyse verbs having selected a verb classification based on the semantic properties of the verbs, used to structure and represent event-related information. In functional grammar, verbs can be categorised in six kinds of processes: material, mental, behavioral, existential, verbal and relational (Halliday, 1994).

According to the frequency results, around 70% of the verbs in both corpora represent material processes, Figures 2a and 2b. However, the verbs included in the material processes category differ in the two corpora. Audio description includes verbs describing motion such as *walk, come, open, fall* etc., which, if separated by the context, do not give explicit information about major events, whereas plot summaries include verbs such as *murder, escape, die, find, help, follow* etc. that refer to the story plot; for example, a murder event may be described in audio description as *he picks up the gun and points at the man…he pulls the trigger*. In plot summaries there are more verbs expressing mental processes (20%) than in audio description (7%). Interestingly, the quality of the mental processes is also different. Mental processes of seeing are mostly depicted in audio description, by verbs such as *watch* and *see*, whereas plot summaries include mental processes related to cognition or affection, what the characters believe and feel, i.e. verbs such as *love, want, know, plan, decide* etc. which are not encountered in audio description.

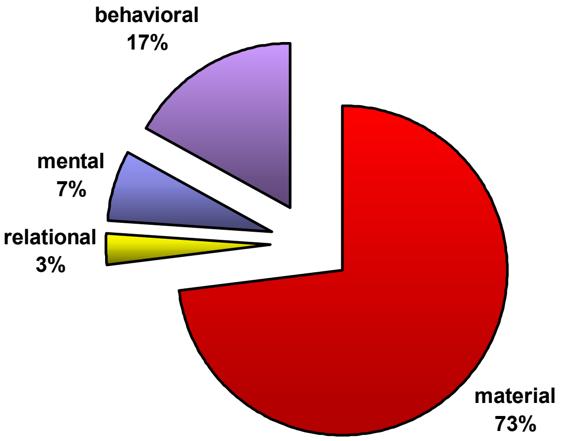


Figure 2a: 4 types of processes in a 376,000-word corpus of audio description based on the 30 most frequent verbs

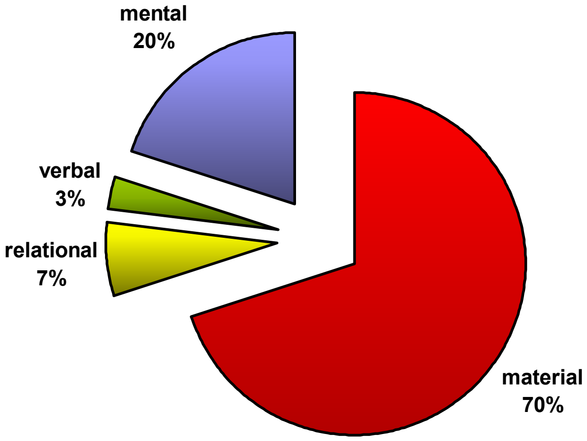


Figure 2b: 4 types of processes in a 9,500-word corpus of plot summaries based on the 30 most frequent verbs

The other verb categories encountered in audio description and plot summaries are different. In audio description, behavioral processes constitute the 17% including verbs such as *smile, stare, look* and *glance*, as the narrators describe what can be seen on screen relatively to the characters physiological and psychological behaviour. These processes may be proved to be important as they can sometimes describe emotions, for example a *laughing* process can express a positive feeling related to the character and concerning the event that has just preceded in the story. On the contrary, the 30 most frequent verbs in plot summaries do not include the behavioral category, as the authors do not describe the character's behavior. Plot summaries also contain verbal processes (3%), such as *tell*, that are not mentioned in audio description due to the dialogue's presence that actually represents the verbal processes.

The frequency results suggest that the same events are described by different verbs in the two corpora. Material verbs may compose the biggest category in both corpora, but the verbs differ completely as shown in Tables 1 and 2.

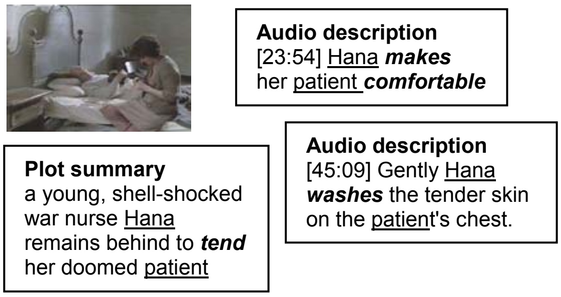
Process	**Verbs in audio description**
Material	open, walk, run, step, hold, close, go, wear, fall, lift, stand, throw, carry, kiss, sit, lead, get, give, cross, join, make, jump
Relational	be
Mental	watch, see
Behavioral	smile, stare, look, glance, nod

Table 1: The 30 most frequent verbs describing 4 types of processes in audio description

Process	**Verbs in plot summaries**
Material	get, love, find, take, kill, help, go, become, plan, die, give, come, escape, make, murder, try, turn, change, follow, lose, need, run
Relational	be, have
Verbal	tell
Mental	want, know, decide, seem

Table 2: The 30 most frequent verbs describing 4 types of processes in plot summaries

In the following example, the *tending* event included in the plot summary is expressed by the verb *tend*, a series of moving images in the film and a series of audio description utterances including the verbal groups *make comfortable* and *wash*, Figure 3. These verbs cannot be matched as they are not synonyms to the verb *tend*.



Audio description
[23:54] Hana ***makes*** her patient ***comfortable***

Plot summary
a young, shell-shocked war nurse Hana remains behind to ***tend*** her doomed patient

Audio description
[45:09] Gently Hana ***washes*** the tender skin on the patient's chest.

Figure 3: Audio description utterances for the same plot summary event

The wordlists of the plot summary and the audio description for the film English Patient do not include any verbs mentioned in both texts. However, they share other open class words; interestingly, the most frequent ones are proper nouns and nouns expressing the characters of the story, locations etc, Table 3.

Common open class words	**OCW PS**	**OCW AD**	**Cumulative OCW**
Hana	1	73	74
Patient	1	33	34
Kip	1	31	32
Caravaggio	1	22	23
Desert	1	17	18
Nurse	1	6	7
Pilot	1	2	3
Burned	1	2	3

Table 3: Common open class words and their occurrence (OCW) in the PS and AD wordlists for the film English Patient

A major event described by one verb in the plot summary, such as *tend* in the example used, may not be explicitly expressed in the audio description, but implied through a series of other events and actions, e.g. *wash* and *make comfortable*. Common event attributes are only the participants *Hana* and *patient*. It is therefore possible to match their combination instead of matching the verb *tend*.

2.2 Creating Test Data

We focus on a method to identify and relate event related information in plot summaries and audio description. The human task involves reading plot summaries and watching the corresponding films, associating the events read to the events visualised. The annotators detect and number the events read in the plot summary. While watching the film, they are told the number of the scene each time a scene commences and they associate the number of the event visualised on screen to the number of the scene, e.g. in the film English Patient, the plot summary event 2 *burned* *horribly in a crash* can be visualised in scene 2 of the film. The human task of matching the events can be characterised as cross-modal event coreference, as humans match events they have read to events they visualise on screen. This had caused disagreements on whether events not explicitely expressed by the visual images but inferred by the sound effects or the dialogue should be annotated or not. The annotation of all events, either explicit or inferred was taken into consideration for the preliminary evaluation of this work due to the multimedia nature of the data included.

2.3 Proposed Algorithms

To compute the human task of event association, we propose a method for cross-document event coreference by identifying and matching verb attributes. The task of event detection in plot summaries has not been automated and main events are already numbered by the human annotators that have read the plot summary. Having identified the main events in the plot summary, we have used the Connexor tagger to represent the plot summary sentences in terms of grammar and functional roles. The algorithms designed generate a list of combinations of event constituents, i.e. verbs and their attributes, according to the tags assigned and match them to the corresponding combinations in the audio description fragments, which are associated with the film data by time-codes and divided into scenes. The scene division was available as part of some scripts by the audio describers who authored the scripts, whereas we have separated the rest of the films according to the scene division in the visual data, i.e. when the location or time changes.

As shown from the verb frequency analysis, in 2.1, it is hard to match verbs from different collateral descriptions expressing the same event. However, characters, plot significant objects and usually locations can be matched. The suggested approach is to match the combination of all or most of the event ingredients, i.e. participants and their roles and circumstances. In the first algorithm, called *Keyword Combination List Generation and Matching (KC)*, the identified plot summary events are grammatically tagged by the Connexor part-of-speech tagger. We then apply rules combining the event constituents, Figure 4; the participants are usually expressed by nouns or proper nouns (as nominal heads), and the circumstances, e.g. location, time, expressed by nouns or adverbs etc. An obligation is to retrieve the combination of the event participants, or one participant and another keyword.

Find Proper Noun / Noun + other keyword:
a. Proper Noun / Noun + Proper Noun-s / Noun-s (+ Noun-s +/ Verb +/ Adverb +/ Adj.)
If no other Proper Noun / Noun is found then find
b. Proper Noun / Noun + Verb +/ Adverb +/ Adj.

Figure 4: A Keyword combination rule

In the sentence *A young, shell-shocked war nurse Hana remains behind to tend her doomed patient*, the algorithm looks for the following combinations: Hana + nurse / patient (+remains +/ tend +/ behind +/ young +/ shell-shocked +/ doomed), as *Hana* is a proper noun and *nurse* and *patient* nouns, and then for the verbs *remains* and *tend*, the adverb *behind* and the adjectives *young*, *shell-shocked* and *doomed*. The next step is to match the generated list of keywords to the audio description utterances including all possible combinations of these keywords without tagging the audio description.

The second algorithm, called *Keyword and Keyword Role Combination List Generation and Matching (KKRC)*, is based on the combination of the keywords and their functional roles in the sentence. Here we have used the machinese syntax function of the Connexor tagger, which assigns words with the roles of subject, agent, object etc. This time, the algorithm looks for the combination of the keywords in the specific roles assigned by the tagger, which means we have to tag the audio description script as well as the plot summary. An example of keyword role combination list rules is shown in Figure 5:

Find [keyword+subject/agent–role] + [other keyword+functional role]:
a.Find [keyword+ subject/agent-role] + [keyword + object-role]
If no [keyword +object-role] is found then
b.Find [keyword+subject/agent-role]+ [keyword + prepositional complement]…

Figure 5: A keyword-role combination rule

In our example, the algorithm generates and matches the combination of *patient* plus the role of object plus another participant, *Hana* plus the role of subject (plus the verb *tend*); Hana[subject] + patient[object] (+tend [verb] etc.).

3 Preliminary Evaluation

The preliminary evaluation of the algorithms has been realised for four films, based on the comparison with human annotations, in terms of precision and recall. We first compare the scenes' identification number of the *Computer-Retrieved Scenes (CRS)* with the scenes' identification number of the *Human Annotated Scenes (HAS)* to find the number of *Correct Computer-Retrieved*

Scenes (CCRS). To find the percentage of the algorithms' precision, we multiply CCRS by one hundred and then divide it to CRS: CCRS + 100/ CRS. To find the percentage of the algorithms' recall, we multiply CCRS by one hundred and divide it to HAS: CCRS * 100/HAS. We have assumed a linear relation between plot summary and film time for the baseline algorithm, which divides the number of the audio description scenes to the number of the plot summary sentences and allocates the first plot summary sentence to the first audio description scene etc. The baseline's low performance (Table 4) is mainly due to the fact that events are ordered differently in plot summaries and in audio description. Film content can be organised in shots and scenes, which relate to film time and the events that comprise the semantic video content, which relate to story time; audio description is temporally aligned with the video data in film time, whereas plot summary is not, relating only to the story time (Salway and Tomadaki, 2002).

Algorithm	Precision	Recall
Baseline	0.1875	0.0261
KC	0.5625	0.6806
KKRC	0.6497	0.4145

Table 4: The evaluation of the algorithms in terms of precision and recall

The evaluation of the KC algorithm presents a significantly better precision and recall than the baseline algorithm. Combining nouns and proper nouns can be useful to find characters although they may not always be plot significant, in which case the precision is low. The KKRC algorithm is more precise, as more retrieved scenes were accurate. Less scenes were retrieved, as assigning roles to characters can be strict sometimes.

4 Discussion

The corpora analysis suggests the heterogeneity of the audio description and plot summaries corpora and the challenge of relating pairs that describe the same events using different verbs, structures and amount of event-related information. This investigation guided the algorithms' approach to match verb attributes; characters and roles, objects, locations or other circumstances. This can show different relations in cross-document structures. The preliminary evaluation shows that precision is of more importance in our case and that semantic role matching is more precise than matching grammatical attributes. To increase the precision, an event classification for filmed stories may be proved useful; for example, the verbs *kill*, *love*, *escape*, *help*, *murder*, *plan* etc. are amongst the 30 most frequent verbs in the plot summary corpus. A preliminary evaluation of using systems such as CYC and WordNet to match events by query expansion has shown that the difference in the vocabulary choice used in the two corpora is not based on synonyms. Matching verb attributes in audio description and plot summaries may also automate the task of event decomposition into other events; for example a *tending* event may include *making comfortable*, *washing* etc. or a *fighting* event may include *kicking, punching, firing at* etc. The algorithms should also be tested on other kinds of data, such as news stories or witness accounts.

5 Acknowledgements

Our thanks to the EPSRC funded project TIWO (GR/R671940/1), the RNIB and ITFC for providing the audio description for this research

References

A. Bagga, A. and B. Baldwin. 1999. *Cross-Document Event Coreference: Annotations, Experiments, and Observations*. Workshop on Coreference and its Applications (ACL99)

A.J. Salway and E. Tomadaki, 2002. *Temporal Information in Collateral Texts for Indexing Video* Procs. LREC Workshop on Annotation Standards for Temporal Information in Natural Language

Connexor: http://www.connexor.com/demo/tagger/

D.R. Radev, and K.R. McKowen. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Journal of Computational Linguistics* 24(3): 469-500

Internet Movie Database: http://www.imdb.com

J. Kuper, H. Saggion, H. Cunningham, T. Declerck, F. de Jong, D. Reidsma, Y. Wilks, and P. Wittenburg. 2003. *Intelligent multimedia indexing and retrieval through multi-source information extraction and merging*. 18th International Joint Conference of Artificial Intelligence, Acapulco

M.A.K Halliday. 1994. *Introduction to Functional Grammar*. 2nd Edition, London

R.K. Srihari. 1995. Computational Models for Integrating Linguistic and Visual Information: A Survey. *Artificial Intelligence Review* 8 (5-6), pp. 349-369

Z. Zhang, J. Otterbacher and D. Radev, 2003. *Learning Crossdocument Structural Relationships using Boosting* 12th International Conference on Knowledge Management

Event Properties: Investigating speaker-generated semantic features for verbs in English and Italian

David VINSON
Dept. of Psychology,
University College London
26 Bedford Way
London, WC1H 0AP, United Kingdom
d.vinson@ucl.ac.uk

Lotte METEYARD,
Gabriella VIGLIOCCO
Dept. of Psychology, UCL
26 Bedford Way
London, WC1H 0AP, United Kingdom
l.meteyard or g.vigliocco @ucl.ac.uk

Abstract

We present analyses of speaker-generated semantic features for words referring to actions and events in English and Italian, assessing the extent to which such features for verbs provide information concerning aspects of event structure, properties of motion and thematic roles of the underlying events. We also investigate how these properties differ across the test languages, focusing particularly upon motion events whose typical expression differs in English and Italian.

1 Introduction

There is a long tradition within cognitive psychology and cognitive neuroscience in using speaker-generated features as a window into the conceptual representation of words. Previous work has revealed the utility of this approach in the domain of nouns referring to objects. For nouns referring to objects, models developed on the basis of speaker-generated features have been used to account for impaired performance in brain-damaged populations (e.g. Cree & McRae, 2003; Garrard, Lambon Ralph, Hodges & Patterson, 2001; Rogers, et al., 2004; Vinson, Vigliocco, Cappa & Siri, 2003). These models have also been shown to be successful at predicting behavioural effects related to semantic similarity (e.g. McRae, de Sa & Seidenberg, 1997; Vigliocco, Vinson, Lewis & Garrett, 2004).

With regard to verbs, although far less work has been performed, previous results illustrate that a model of lexical-semantic representation developed on the basis of speaker-generated features (FUSS: Featural and Unitary Semantic Space, Vigliocco, et al., 2004) can successfully predict semantic similarity effects, just like for the nouns referring to objects (Vigliocco, et al., 2004; Vigliocco, Vinson, Damian & Levelt, 2002).

Given the ability of our model based on speaker-generated features to predict semantic similarity effect, in the current work we explore whether and to what extent properties of events, as discussed in linguistic proposals (described in more detail below), are captured by specific features in our database.

The database we consider includes 216 English verbs and 216 Italian verbs (Vinson & Vigliocco, 2002; Sbisà, Collina, Vinson, Tabossi, & Vigliocco, 2003). For each word, 20 participants generated semantic features; each feature has an associated "feature weight" corresponding to the number of participants who produced that feature for a given word; feature weights correspond to the salience of a particular feature to that word's meaning.

The particular semantic fields we chose to focus upon were body actions (e.g., bleed, tickle), change of location (e.g., drag, push, pull), change of state (e.g., blend, mix, stir), communication (e.g., preach, scream, talk), construction (e.g., build, make, repair), contact (e.g., bump, hit, press), cooking (e.g., bake, roast, steam), destruction (e.g., break, destroy, smash), exchange (e.g., borrow, donate, lend), light emission (e.g., flicker, glow, sparkle), direction of motion (e.g., arrive, descend, enter, rise), manner of motion (e.g., creep, limp, stagger), noises (e.g., chime, rattle, screech), animal noises (e.g., oink, meow), and tool actions (e.g., chop, cut, hammer, saw). We consider the distribution across semantic fields of feature classifications of three broad types: features related to properties of event structure, features related to properties of motion events, and features related to thematic roles.

In order to investigate cross-linguistic differences, we focus particularly upon the contrast between verbs referring to direction and manner of motion, whose semantic content may differ between English and Italian. These languages differ in the information typically contained in verbs describing motion (Talmy, 1985): in English verbs are typically "satellite-framed", expressing manner in the verb (e.g. stagger, amble, shamble, etc.) and path through verb particles (e.g. "into, "across", etc.), while Italian motion verbs are typically verb-framed, expressing path of the motion in the verb (e.g. entrare, to enter; uscire, to exit),and manner through adverbial modifiers (e.g. correndo, running). Through featural analyses we investigate whether this general difference between verbs in the two languages results in differences in speaker-generated features.

2 Event Structure

We assess two properties of events, telicity (reflecting an endpoint) and durativity (reflecting an interval), distinctions which allow classification of events into four types (Vendler, 1967; Dowty, 1979), as in Table 1:

	- telic	+ telic
- durative	STATE	ACHIEVEMENT
+ durative	ACTIVITY	ACCOMPLISH-MENT

Table 1: Classification of event types

Properties of features can be used to illustrate how verbs in different semantic fields reflect different types of events: Features related to duration should reflect preference along the durative dimension; features reflecting a change of state should reflect preference along the telicity dimension.

2.1 Features reflecting duration

We defined features related to duration as those overtly reflecting time period or speed of activity (e.g. <brief>[1], <constant>, <fast>, <short-duration>), also reflecting periodicity (e.g. <daily>, <repeated>). We excluded features which might express duration but required additional inference (e.g., <flash>, <hold>, <process>).

For each word, we summed the weights of all features classified as reflecting duration, and divided that total by the summed weight of all features generated for that word; this figure reflects the relative importance of duration features in a given word (this same procedure was followed for all feature type classifications).

Considering the distribution of duration features across semantic fields (averaging the proportion of duration features across all words in a given semantic field, then averaging across languages), the fields with the most duration features overall were manner of motion (4.1%), light emission (2.4%) and noises (2.3%); fields such as change of location, communication, cooking, construction and destruction all had less than 1% of duration features.

In English, the words containing the highest proportion of duration features included mostly verbs of motion: stay (28%), run (22%), jog (21%), lend (13%); the highest Italian words were correre (run, 13%), lampeggiare (flash, 12%), tintinnare (tinkle, 11%) and rimbalzare (bounce, 10%).

We assessed cross-linguistic differences in the production of duration features by examining the difference between English and Italian production frequencies for each semantic field, dividing each by the overall rate of production of duration features in that language (1.7% across all words in English; 1.3% in Italian); this same method was used for all subsequent feature classifications as well. The largest resulting relative differences occurred for the fields of light emission (1.2% in English, 3.6% in Italian), body actions (2.9% in English, 1.4% in Italian) and contact (1.2% in English, 2.3% in Italian). English speakers were broadly more likely to produce duration features for movement involving the body, while Italian speakers were more likely to produce duration features for events involving inanimate entities.

Concerning verbs referring to manner of motion and direction of motion, speakers of both languages were similar in production of duration features, producing many more for manner of motion (4.1%) than for direction of motion (1.2). This difference is consistent with the distinction of such words along the durative dimension: manner of motion verbs typically express activities, while many direction verbs typically express achievements (instantaneous transitions into a goal state, such as “enter” and “arrive”).[2]

2.2 Features reflecting change of state

Features related to change of state included not only overt change of state features (e.g. <change>, <consume>, <destroy>), but also start/end points (e.g. <begin>, <end>), and instances of exchange (e.g. <exchange>, <give>, <get>) which reflect change of state regarding ownership.

Considering semantic fields, the fields with the most change of state features across both languages were destruction (33%), construction and exchange (22%), change of state (21%) and cooking (17%); the fewest such features were produced for animal noises (2%), communication (3%), manner and direction of motion (both 5%).

[1] Examples of speaker-generated features are presented in English throughout; the same definitions were always used to classify features in both English and Italian.

[2] Most often, event classification requires more sentence information than the verb alone□ these results nonetheless illustrate differential typical behaviour between manner and direction motion verbs.

This generally is consistent with the distinction between telic and atelic events: those semantic fields with the highest proportion of change of state features are also the most likely to express telic events.

The English words containing the highest proportion of change of state features included exchange (58%), break (55%), blend, make, destroy (51%); the highest in Italian were fracassare (smash, 34%), scambiare (exchange, 34%), regalare (give as a gift, 31%), distruggere (destroy, 30%).

As for duration features, English speakers produced more change of state features than Italian speakers (13.3% vs. 8.2%). At the semantic field level, the greatest difference was observed for words referring to construction (English 34.5%, Italian 9.3%). Importantly, the pattern of results for change of state features also differed across languages of direction and manner of motion: Italian speakers produced more change of state features for direction of motion than for manner of motion (4.9% vs. 4.3% respectively), while the pattern was reversed for English speakers (4.5% vs. 6.2%).

This latter difference between English and Italian is consistent with the relative importance of manner and path to verb representations in each language: manner is generally more important in English verbs while path is generally more important in Italian verbs. This correspondence may render events from these particular semantic fields (English manner of motion verbs and Italian direction of motion verbs) more salient as events with consequences.

3 Motion events

Languages differ in the manner in which events involving motion are expressed in sentences (Talmy, 1985); they vary in the expression of motion path, manner, figure (the object in motion) and ground (the reference object). Here we investigate the extent to which specific differences between English and Italian are reflected in speaker-generated features, as discussed in the Introduction. Given the opposing tendencies in which manner is coded in the verb itself more frequently in English, and direction in Italian, we would expect features to differ concordantly since they are a reflection of the salient properties of words. We examine features of four types for this purpose: those depicting the path of motion, manner of motion, the figure of motion and the ground of motion. These analyses were applied only to verbs of motion (direction of motion vs. manner of motion).

3.1 Features reflecting path of motion

We defined features related to path of motion as those depicting direction (e.g. <back>, <back-and-forth>, <down>, <fall>, <forward>), including relative motion (e.g. <direction-in>, <leave>). This classification also included exchange type features which reflect (abstract) direction of motion (e.g. <give>, <receive>), but does not include simple presence/absence of motion in which path is not actually expressed (e.g. <move>, <motionless>, <remain>).

In both languages, the words containing the highest proportion of path features referred to direction of motion. In English the most path features were produced for enter (62%), return, descend (58%), arrive and ascend (47%); in Italian, for discendere (descend, 47%), entrare (enter, 46%), venire (come, 38%) and salire (go up, 33%). No overall difference was observed between English and Italian, in both languages there were substantially more path features for verbs referring to direction of motion (31.3%) than for those referring to manner of motion (7.5%). This correspondence reveals that verbs which specifically encode direction (as indicated by our intuitive classification into manner/direction verbs) commonly elicit path-related features, despite the difference in typical expression of verbs in the two languages.

3.2 Features reflecting manner of motion

Features related to manner of motion were those overtly reflecting manner (e.g. <awkward>, <curve>, <drag>, <fast>, <out-of-control>), those reflecting means of transportation (e.g. <airplane>, <car>), and those related to orientation of a moving entity (e.g. <low>, <upright>). This classification did not include body-parts by which motion is achieved (<uses-foot>, <uses-leg>, etc.).

In English, the most manner features were produced for wander (48%), stagger (46%), creep (43%), jog (41%); for Italian the most were produced for marciare (march, 41%), vagare (wander, 39%), zoppiacare (limp, 33%) and saltellare (skip, 28%).

Speakers of the two languages were highly consistent in producing manner features for verbs we had judged to depict manner of motion (18.9%) than for those depicting direction of motion (8.5%). As was the case for path of motion features, speakers of both languages generated manner-related features for verbs overtly depicting manner of the motion, even when this is contrary to the overall tendencies of the language.

3.3 Features reflecting figure of motion

Features related to the figure of the motion reflected the entity that is moving in any motion event. These can overtly depict the figure itself (e.g. <ball>, <by-human>, <vehicle>), but also parts of the figure (e.g. <wheel>). For this reason any event involving motion of any kind is considered (e.g. <face> for actions like sneeze, <blink>). The feature <by-humans> was considered as "figure of motion" whenever it reflected motion by a human entity.

In English the most figure features were produced for send (43%), drive (42%), throw and pedal (39%); in Italian the most figure features were produced for volare (fly, 29%), inseguire (pursue, 28%), affondare (sink, 16%) and seguire (follow, 14%).

Overall, manner of motion had more figure features (18.9%) than did direction of motion (7.5%). Unlike features referring to manner and path, however, we observed an interesting cross-linguistic difference (despite overall differences in the base rate of such features across languages): English speakers produced the most figure features for manner of motion (19.8%), with very few such features for direction of motion (7.7%); the pattern reversed for Italian speakers with substantially more figure features for direction of motion (7.3%) than for verbs referring to manner of motion (4.1%).

Here, the cross-linguistic distinction in the way in which manner and path are expressed in verbs is borne out in the features. For each language the motion verbs with the aspect most typically coded in the verb (manner or path) have more motion features produced for them which refer to the mover. This is consistent with the salience effect for change of state features (section 2.2), in which figures of motion are more salient for those types of verbs most common in a language, and less salient for atypical types (direction verbs in English; manner verbs in Italian).

3.4 Features reflecting ground of motion

Features related to the ground of the motion reflected a non-moving entity upon which action occurs (e.g. <church>, <floor>, <ground>, <kitchen>.Building materials are not considered to be ground (e.g. <wood> for "to hammer). The feature <by-humans> was considered as "ground of motion" whenever it reflected a non-moving human entity upon whom an action can be considered to occur (e.g. "to give", "to receive", "to punch"). Most <by-human> features are also coded as "figure", as they can also reflect a human in motion.

By semantic field, ground features were most common for direction of motion (11.8%), manner of motion (6.7%) and least common for change of location (4.9%). In English, ground features were produced most often for put (35%), wade (27%), dive (23%) and fly (19%); Italian speakers produced the most ground features for entrare (enter, 38%), affondare (sink, 27%), tuffarsi (dive, 27%) and nuotare (swim, 25%).

For manner of motion, features referring to ground of motion were more common in English (7.5%) than Italian (5.9%), while this pattern was reversed for words referring to direction of motion (9.5% in English; 13.9% in Italian). As for figure of motion and change of state features, ground-related features were most commonly produced for those verbs that reflect the typical expression of motion events in a particular language.

4 Thematic roles

This area of investigation considers the different kinds of entities that can participate in events depicted by a given verb, assessing claims that thematic roles can be considered to be verb-specific and reflected in featural distribution (McRae, Ferretti, Amyote, 1997). Here we investigate the extent to which speaker-generated features reflect typical thematic roles for different verbs; including agent, patient/theme, instrument, and location, across all semantic fields in the database of speaker-generated features.

4.1 Features reflecting Agents

Features related to Agent depict a sentient causer of the action (e.g. <by-humans>, <by-adult>, <by-animal>, <carpenter>). This classification also includes features which depict awareness/sentience of the actor (e.g. <intelligent>, <intentional>, <desire>, <want>).

By semantic fields Agent features were most common for animal noises (28%, nearly all of which reflected the animal that makes a particular sound), communication (16%) and exchange (11%); the fewest were produced for light emission (1%), cooking (2%), tool action, noise and change of state (5%). In English, Agent features were most common for want (57%), notice (42%), oink (41%), meow (37%); Italian speakers produced Agent features most often for grugnire (oink, 30%), consigliare (suggest, 26%), marciare (march, 25%) and miagolare (meow, 23%).

No particular cross-linguistic differences were observed in this classification. Concerning motion verbs, speakers of both languages produced Agent features approximately equally often for manner of motion verbs (8.9%) and direction of motion verbs (7.6%), reflecting the fact that the causers of motion verbs are comparable in both languages, despite the differences in whether manner or direction is typically expressed across languages.

4.2 Features reflecting Patient/Theme

Features related to Patient/Theme included properties or identification of the entity most affected by the action (e.g. <army> "to command", <baseball> "to hit", <clothing> "to borrow"). Also included were features related to consequences of the action upon the patient/theme (e.g. <change location> for "to carry", <change shape> for "to bend", <death> for "to kill"). Agents were excluded from this classification (except in features like <2-humans> which can depict both agent and patient). Features which apply only to a subset of possible patient/themes for a given action (e.g. <heavy> for "to carry", which does not apply to everything that can be carried) were still classified as "related to patient/theme".

By semantic fields patient/theme features were most common for cooking (38%), construction and change of state (both 35%), exchange and contact (both 28%), and least common for animal noises (1%), manner of motion (8%), direction of motion (10%) and noise (11%). In English, patient/theme features were most common for burn (75%), steam (66%), blend, pour (63%) and repair (62%); in Italian the most patient/theme features were produced for discutere (discuss, 40%), riempire (fill, 39%), prendere (take, 36%) and scambiare (exchange, 34%).

Features of this type were produced substantially more often by English speakers than by Italian speakers overall (23% and 13% respectively); beyond this difference, Italian speakers produced more patient/theme features for communication verbs (English 8.5%, Italian 17.6%), and English speakers produced more for cooking verbs (English 54.3%, Italian 21.0%), which may simply reflect cultural differences in attitudes toward cooking.

Considering motion verbs, the two languages exhibited similar performance: slightly more patient features were produced for direction of motion verbs (9.8%) than for manner of direction verbs (8.4%). This may reflect differences in event typology: the tendency for manner verbs to depict activities (which often do not include patients) and direction verbs to depict achievements and accomplishments (both of which typically include patients; see section 2.1). Because information about the specific event type is often expressed outside sentences' main verbs in both languages, it is not surprising to find common patterns of patient/theme features for motion verbs.

4.3 Features reflecting Instruments

Features related to Instruments overtly reflected the instrument by which the action is achieved, including body parts where appropriate (e.g. <use-arm>, <use-hammer>, <machine>, <use-tool>). This classification also included properties of the instruments by which the action is performed (e.g. <electric>, <metal>, <sharp>).

By semantic fields instrument features were most common for cooking (24%), tool action (21%), construction and contact (both 16%); fewest were produced for direction of motion (4%), exchange (5%), animal noises and light emission (both 6%), and change of location (7%). In English instrument features were produced most often for write, steam (both 43%), bake (42%), smoke and drive (both 38%); in Italian they were most common for calciare (kick, 38%), pedalare (pedal, 35%), scrivere (write, 34%) and disegnare (draw, 31%).

Beyond the base difference between production of Instrument features in English and Italian (12.7% in English, 10.3% in Italian), more extreme cross-linguistic differences were found for Instrument features for words referring to cooking (35% in English, 13% in Italian), and also for contact (11% in English, 20% in Italian).

For motion events, both languages were similar in that substantially more instrument features were produced for manner of motion events (11.9%) than for direction events (3.9%); most of these features reflected more salient use of body parts (as instruments) to carry out manner of motion verbs, than to carry out direction of motion verbs (for which instead patient/theme features were more salient; see section 4.2).

4.4 Features reflecting Locations

Features related to Location depict where an action occurs. This classification largely overlaps with Ground (section 3), but not completely (for example, where Ground is a non-moving human entity); examples include <church>, <garden>, etc.

By semantic fields location features were most common for direction of motion (15%), manner of motion (8%), change of location (6%) and cooking (5%); hardly any location features were produced for verbs of destruction, light emission, noises, animal noises, communication, construction and contact (all less than 1%). In English location features were most common for arrive (40%), put (37%), go (29%), leave (28%) and wade (27%); in Italian they were most common for entrare (enter, 36%), affondare (sink, 28%), tuffarsi (dive, 27%) and nuotare (swim, 25%).

Speakers of both languages produced location features fairly infrequently (3.6% in English, 2.8%

in Italian), nonetheless some differences were observed. English speakers were much more likely to produce location features for verbs related to cooking (9.0%, vs. only 1.9% for Italian).

For motion verbs, location features were produced much more often in both languages for direction of motion verbs (14.9%) than for manner of motion verbs (7.5%); Like all the other thematic roles investigated here, no language differences were observed in this classification.

5 Conclusion

Through the use of speaker-generated features, we found substantial commonalities between properties of verbs in English and Italian. Besides finding broad overall convergence across semantic fields, we also observed specific relationships between theoretical constructs and feature typology. Concerning event typology, this was most evident for features related to duration, which were in broad correspondence with typical event types expressed with verbs in different semantic fields. Concerning thematic roles, speakers of both languages produced broadly common types of features for verbs from a variety of semantic fields, illustrating aspects of agents, patients/themes, instruments and locations that have differing relevance for words from different semantic fields.

We also observed numerous differences between English and Italian with respect to the salient information encoded in verbs related to motion events, information which typically differs across the two languages. This difference had broad consequences. First, change of state features were more commonly produced by English speakers for manner verbs, but by Italian speakers for direction verbs, reflecting a difference in salience that is presumably related to the way in which events are typically expressed in the languages. A similar pattern was observed for features related to Figure and Ground: more such features were produced by English speakers for manner verbs, and by Italian speakers for direction verbs. Nonetheless, speakers of both languages were consistent in producing features related to manner and path of motion for verbs that expressed that information, even when inconsistent with the overall tendencies in their language. These results highlight the utility of speaker-generated features in illuminating specific questions about cross-linguistic similarity and variation, and illustrate how cross-linguistic differences can have far-reaching semantic consequences.

6 Acknowledgements

This work was supported by Human Frontier Science Program grant (HFSP148/2000) and by Economic and Social Research Council grant (RES000230038) to Gabriella Vigliocco.

References

Cree, G.S. & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132, 163-201.

Dowty, D. R. (1979). *Word meaning and Montague Grammar*. Dordrecht: D. Reidel

Garrard, P., Lambon Ralph, M.A., Hodges, J.R., & Patterson, K. (2001). Prototypicality, distinctiveness and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18, 125-174.

McRae, K., de Sa, V. & Seidenberg, M.C. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126, 99-130.

McRae, K., Ferretti, T., and Amyote L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12, 137-176.

Rogers, T. T., Lambon Ralph, M. A, Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, 111, 205-235.

Sbisà, S., Collina, S., Vinson, D.P., Tabossi, P. & Vigliocco, G. (2003). Feature-based similarity measures for 460 Italian words [unpublished data]

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description, vol. 3: Grammatical categories and the lexicon.* Cambridge: Cambridge University Press

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

Vigliocco, G., Vinson, D.P., Damian, M.F. & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition*, 85, B61-B69.

Vigliocco, G., Vinson, D.P, Lewis, W. & Garrett, M.F. (2004). The meanings of object and action words. *Cognitive Psychology*, 48, 422-488.

Vinson, D. P., Vigliocco, G., Cappa, S.F. & Siri, S. (2003). The breakdown of semantic knowledge along semantic field boundaries: Insights from an empirically-driven statistical model of meaning representation. *Brain and Language*, 86, 347-365.

Morphological priming with French verbs:
the neglected role of orthographic neighbourhood.

Madeleine VOGA-REDLINGER
Université de Provence
Personal Address :
26, rue de l'Aiguillerie,
34000 Montpellier, France
redlinger@polytech.univ-montp2.fr

Hélène GIRAUDO
Laboratoire Parole et Langage, (LPL),
CNRS et Université de Provence
giraudo@up.univ-aix.fr

Abstract

The present study used the masked priming paradigm to examine effects of morphologically related primes on the recognition of infinitive forms of French regular verbs, while varying the level of neighbourhood density. Morphological effects were assessed relative to form-related control primes, while an incremental priming technique was used . Neighbourhood density was found to affect the presence or absence of morphological facilitation, since high neighbourhood density inflections failed to prime their base form, contrariwise to low neighbourhood size verb category. These results are interpreted within an interactive activation model of morphological representation, and more importantly, they put forward a possible explanation for some discrepant results in the literature.

1 Introduction

Since Rumelhart & McClelland (1986) first presented their connectionist model of the English past tense system, inflectional morphology has acquired a particular significance in debates about the nature of cognition. The representation and processing of the English verb has become the battleground of two approaches: one that postulates that mental grammar is used directly for language processing and that the processing system makes the same distinctions as grammar does (Miller & Chomsky, 1963; Jackendoff, 1997) and one that postulates the total absence of the regular-irregular distinction in the processing system. According to the latter (Plaut & Gonnerman, 2000) the role of morphology is captured by a general processing system that treats both regular and irregular forms in the same manner, in such a way that there is no need to hypothesise the explicit representation of morphology in the mental lexicon; besides, there is no mental lexicon in such a model.

Extensive research on the controversy between symbolic (Pinker & Prince, 1988) and connectionist accounts of the acquisition and processing of the English past tense and of verb morphology in general has effectively reached stalemate as far as the observable properties of the process are concerned. Current research is directed to examining the properties of the neural systems that underlie processing and representation of linguistic material (Marslen-Wilson & Tyler, 1998).

Research has also turned to other languages, morphologically richer than English, such as Greek (Tsapkini, Jarema & Kehayia, 2002) or Hebrew (Frost, Deutsch & Forster, 2000), in order to definite the nature of cognitive mechanisms implicated in the lexical access of verbs and morphologically complex words in general. Some researchers (Tsapkini, Jarema & Kehayia, 2002; Pastizzo & Feldman, 2002; Voga & Grainger, 2004) question the relevance of such distinctions as regular-irregular, at least as far as the masked priming technique is concerned, without necessarily postulating the absence of morphological representation in the lexicon. There is a growing body of evidence that the dichotomy between regular and irregular inflections may not be as sharp as claimed by the symbolic account. This evidence does not only come from highly inflected languages like Greek (Tsapkini et al., 2002, Voga & Grainger, 2004), but also from English data : indeed Pastizzo & Feldman (2002) find that relatively to an orthographic control (as opposed to the standard unrelated baseline) irregular past inflections like "FELL" prime their base form (33ms) "fall" almost as strongly as regular forms (e.g. HATCHED – hatch, 44ms of facilitation), while irregular inflections of diminished orthographic similarity do not induce significant priming (e.g. "TAUGHT-teach", 15ms). The reason why the regular-irregular issue is central to our reasoning is that it shapes the modelling of the morphological processing of languages other than English, and inflections other than the irregular past tense. For example, one important issue addresses the way French irregular inflections (present as well as past) are processed, and this issue cannot be isolated from the dual-route versus one mechanism controversy.

At the same time when the regular-irregular debate was monopolising a great amount of scientific interest, great progress has been made on the representation and processing of orthographically / phonologically related items in the mental lexicon. With the masked priming technique, evidence was found that orthographic similarity of the prime affects (inhibits) lexical access of morphologically complex targets, despite (or because of) the absence of any morphological relation between them (Grainger, Colé et Ségui, 1991). For example, the prime "mûrir" inhibits the target "MURAL" and this inhibition attains 27ms for words that share their initial letters. This inhibition is accounted for in terms of "preactivation of lexical representations during the processing of the prime, that interfere with the processing of the target" (Grainger, Colé & Ségui, 1991, p.380). The inhibitory effect of a prime like "blue" on the target "BLUR" (Ségui & Grainger, 1990) is found, according to the same logic, because "blue" is a very powerful competitor in the recognition process of its neighbour "BLUR". The presentation of "blue" as a prime does nothing less than reinforce its competitor status, already quite important (because of its frequency), thus delaying target processing. This inhibition of O+M- primes combined with the absence of such an effect for nonword primes, is also found by Drews and Zwitserlood (1995) on derivational morphology in German and Dutch. The fact that nonword primes do not behave in the same manner argues in favour of the hypothesis that this competition indeed does take place in the lexical level.

Within such a dynamic interactive activation architecture of visual word recognition, it can be argued then that the recognition of a verb will be driven by the extraction of its linguistic features in a way that the visual presentation of the stimulus word at the entry of the cognitive system (prime) will produce the successive activations of all its characteristics at different (interconnected) levels of processing. Because many words share multiple features (phonological, orthographic, morphological and semantic), word recognition implies, in the very early stages of processing, multiple competitions between word candidates that could match a given stimulus. For instance, if we consider the French word "*mentons*" (we lie), it can potentially activate "*mentir*" (to lie) as well as "*sentir*" (to smell) or "*sentons*" (we smell) or "*menton*"(chin), because these word representations share either formal or semantic features or both, in which case they share a common stem morpheme. Thus processing of morphologically complex forms does not simply consist in activating its lexical representation. The processing system has to make the right "choice" as to which candidate should be activated the most (or the first).

The experiment reported here is designed to test the role of orthographic neighbourhood on the recognition of French verbs. If the rationale exposed above corresponds to the way things happen in the processing system, then verbs which have fewer neighbours (irrespectively of the grammatical category of these neighbours) should produce more morphological priming than verbs that have more neighbours, i.e. that share formal features with many other words. We used the masked priming paradigm, developed by Forster and Davis (1984), in which the prime is presented for a duration that does not permit conscious identification. This technique helps avoiding any strategic processing based on the controlled relations between prime and target, and sheds light on the automatic cognitive processes governing lexical access.

2 Experiment 1

2.1 Method

2.1.1 Participants

Thirty subjects who reported normal or corrected-to-normal vision participated in the experiment. They were first and second year Psychology students from the University of Aix-en-Provence. Participation was rewarded with **some** extra points for Psychology courses.

2.1.2 Stimuli and design

Fifty-four French words and fifty-four nonwords were used as targets. Targets were always the infinitive form of French verbs, all regulars, 4 to 9 letters long (mean: 5.5 letters) with an average frequency of 69.40 occurrences per million (New, Pallier, Ferrand, & Matos, 2001) and consisted of 1) 27 verbs that had a large number of orthographic neighbours, and 2) 27 verbs that had a neighbourhood density as low as possible. These two categories of target word represent the two levels of the orthographic neighbourhood factor. The criteria used to decide if a word was or was not an orthographic neighbour of a verb were not as strict as those used for nouns, simply because there are less verbs than nouns in the language. Words were considered as neighbours of a verb if they shared the letters that form the root of this verb. In the French language, verbs have a characteristic ending, which can be "-er", "-ir", "-

re”, or “-oir” to cover most of the cases. Following this criterion, a verb like “porter”, where the root is “port” and the ending is “er” has numerous neighbors, like “portail” (“portal”), “porte” (“door”), “port” (“harbour”), “portier” (“porter”), “portion” (“portion”), “portique” (“porch”), “portrait” (“portrait”), “portière” (“door”), “portugais” (“portuguese”) etc. A verb like “mourir”, on the other hand, has a very small neighbourhood, limited to the rare “mouron” (“scarlet pimpernel”). The number of neighbors was estimated with the help of **a** French dictionary (Petit Robert).

Each target was given three types of prime: a repetition prime, a morphologically related prime, and an orthographically related prime (see Table 1 for examples). Targets were always presented in lowercase letters and primes in uppercase letters, in order to minimize the visual overlap between prime and target, and stress markers were preserved in lowercase letters.

Morphologically related primes were inflections of the infinitive (base) form: there were 15 past participles, 26 present tense forms, 4 future forms as well as 9 “imparfait” (past continuous) forms. Orthographically related primes shared a high proportion of letters with targets, but had no semantic or morphological relation with them. The orthographically related primes were used as the control baseline. In fact, like Giraudo & Grainger (2003) demonstrated (see also Pastizzo & Feldman, 2002), morphological facilitation should be estimated relatively to an orthographically related baseline and not only relatively to the unrelated condition, in order to evacuate form effects, and this is particularly relevant when using priming techniques.

54 French nonverbs were created respecting the phonotactic constraints of the language and were matched for length with the real verbs. The primes for nonword targets matched the word primes in terms of orthographic overlap, and were constructed so as to mimic the morphologically and orthographically related primes for word targets. Three experimental lists were created by rotating targets across the three priming conditions using a Latin-square design, so that each target appeared only once for a given participant and for a given prime duration, but was tested in all priming conditions across participants. Participants were randomly assigned to one of the three lists.

Primes	**Target**	
	monter (climb)	**écrire** (write)
Morphological	**montais** (was climbing)	**écrit** (written)
Orth. overlap	3.48 letters in common	3.74 letters in common
Orthographical	**montagne** (mountain)	**écrin** (jewel-) case
Orth. overlap	3.4 letters in common	3.62 letters in common

Table 1: Examples of Stimuli Employed in the Six Experimental Conditions.

2.1.3 Procedure and apparatus

The experiment was conducted on a PC computer using DMDX (Forster & Forster, 2003). Each trial consisted of three visual events. The first was a forward mask consisting of a row of nine hash marks that appeared for 500ms. The mask was immediately followed by the prime. The prime was in turn immediately followed by the target word which remained on the screen until participants responded. The intertrial interval was 500ms. The three prime durations used in this experiment were 26, 40 and 53ms. As is commonly reported in the literature, these prime durations do not permit conscious identification of the prime (e.g. Pastizzo & Feldman, 2002). Each participant was tested in all durations for a given list, which means that he saw each target 3 times (one time for each duration). Experimental trials were randomized in such a way that an item of a given duration was never followed by an item of the same duration and a given target was never followed by the same target in another duration. We developed a small software tool to carry out this type of randomization, which minimizes repetition effects and responding strategies.

Primes appeared in the middle of the screen presented in Times New Roman lowercase characters (16 point) and targets in uppercase letters. The participants were seated 50 cm from the computer screen. They were requested to make lexical decisions on the targets as quickly and as accurately as possible, by pressing the appropriate key on the computer keyboard. After 16 practice trials, participants received the 324 experimental trials in one block, interrupted by two brief pauses.

2.1.4 Results

Correct RTs were averaged across participants after excluding outliers (RTs>1000ms, <400ms). Two items were excluded from the analysis, because of high error rates. Results are presented in Table 2. An ANOVA was performed on the data with prime type (repetition, morphological, orthographic control), neighbourhood size (high or low) and prime duration (26, 40 or 53ms) as independent variables.

D ms		R	M	O	O-R	O-M
26	HND verbs	540	541	547	7	6
	LND verbs	541	554	552	11	-2
40	HND verbs	527	543	542	15	-1
	LND verbs	534	540	554	20	14
53	HND verbs	532	545	551	19	6
	LND verbs	534	539	553	19	14

Table 2. Reaction times (RT in milliseconds) for lexical decisions to targets in the repetition (R), morphological (M), and orthographic control (O) prime conditions, for the three prime durations (D). Net priming effects are given relative to the orthographic condition.

There was a significant main effect of prime type, F1(2, 58) = 10.64, p<.001, F2(2, 100) = 5.38, p<.01, with targets preceded by an identity prime being responded to faster than those preceded by orthographical control primes. The main effect of neighbourhood size was not significant, F1(1, 29) = 2.13, F2<1, nor was the overall effect of prime duration (both F1<1, F2(2, 100) = 1.56. The interaction between prime type and neighbourhood density was not significant (both Fs<1) nor was the triple interaction (prime type x neighbourhood density x prime duration), F1(4, 116) = 1.18, F2<1. What is of more interest for the hypotheses under study is the partial effects of main factors for the duration of 40 and 53ms. The effect of prime type was significant for the verbs with a small neighbourhood density for the durations of 40ms F1(2, 58) = 6.38, p<.01 and 53ms, F1(2, 58) = 3.98, p<.05, but it was not significant for the verbs who had a big neighbourhood, F1(2, 58) = 2.74 and F1(2, 58) = 2.98 respectively. Planned comparisons revealed that the difference between morphologically related and control conditions was significant for the verbs having a small number of neighbours, for the duration of 40ms, F1(1, 29) = 6.45, p<.05, F2(1, 24) = 1.79, as well as for the duration of 53ms, F1(1, 29) = 4.68, p<.05, F2(1, 24) = 1.64. On the other hand, morphologically related primes of big neighbourhood verbs did not induce significant priming (all Fs<1). This difference between the two categories of targets is statistically significant only by subjects. This critical pattern of effects, that is, morphological facilitation for small neighbourhood verbs and absence of morphological effect for big neighbourhood verbs, is shown in Figure 1.

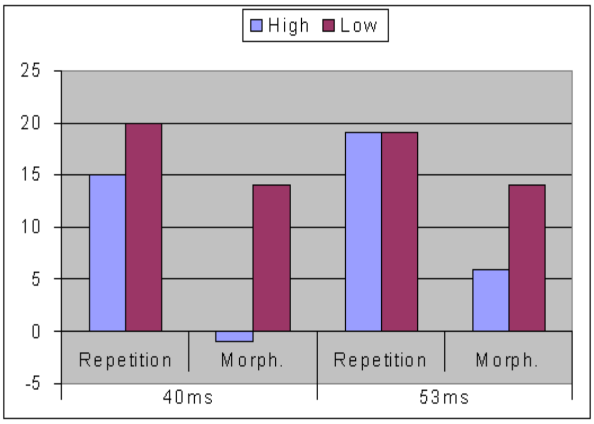


Figure 1: Net priming effects (in ms) for the repetition and the morphological conditions as a function of prime duration (40 and 53ms) and orthographic neighbourhood density (high or low).

3 Discussion

The experiment reported in this paper focused on the orthographic neighbourhood size in the context of verb processing, i.e. the number of orthographic neighbours that a given verb can potentially activate when presented to the visual cognitive system. A word was called “an orthographic neighbour” of another word when it shared initial letters/phonemes (to the same degree as this word shared with its morphologically related prime), but without having any morphological relationship with it. Thus, our measure was the opposite of the morphological family size of a stem, defined as the number of different complex words in which the stem appears as a constituent. This morphological family size has been found to affect response latencies in tasks such as visual lexical decision (Schreuder & Baayen, 1997) and in a variety of languages, Germanic or Semitic (dutch : Schreuder & Baayen, 1997; Bertram, Baayen & Schreuder, 2000; English : De Jong, Feldman, Schreuder, Pastizzo, Baayen, 2002; Hebrew : Moscoso Del Prado Martin, Deutch, Frost, Schreuder, De Jong, Baayen, submitted) and reflects the amount of words that will work as “synagonists”. Our measure of “orthographic neighbours” reflects the amount of words that will work as “antagonists”, at least as far as an interactive activation view of the lexicon is concerned.

We hypothesised that verbs of a larger orthographic neighbourhood will produce morphological priming of a reduced amplitude compared with verbs having fewer neighbours, the later having no (or very few) competitors on the lexical level. Our results confirmed this prediction, since high neighbourhood density verbs failed to induce priming, whereas low neighbourhood density verbs induced significant morphological facilitation (though significant only by subjects). The fact that verbs of both categories induced significant repetition priming demonstrates that primes were indeed correctly processed and that the two categories of verbs do not differ significantly as to their other features. Thus, we can attribute the difference observed in morphological facilitation solely to the manipulated variable.

In an interactive activation dynamic network, where the morphological level is situated above the lexical level (supra-lexical account of morphology, Giraudo & Grainger, 2001) the facilitation that the prime "port-ons" ("we carry") induces to the recognition of the target "port-er" ("to carry") is the result of concurrent activation and inhibition: the activation would come from the shared morphological unit (through feed back from the morphological to the lexical level) and the inhibition would be the result of the intra-level inhibitory connexions at the lexical level. Following this logic, the morphologically related prime of a high neighbourhood density verb will have to resolve the competition from all its neighbours (like "porto" or "porte" or "portrait") before reaching the activation threshold required for the identification of the corresponding representation. A prime having no (or only a few) neighbours will activate the corresponding lexical representation in a more efficient way. Briefly stated, we have shown that when a prime has many orthographic neighbours (sharing no morphological relation with it) it will induce less morphological priming than a prime that has very few neighbours.

Of course, the results we present do not resolve the question of the English past tense, since they concern the processing of the French inflection. Nevertheless, we consider that the variable "orthographic neighbourhood size" has been neglected, although a great amount of research has been accomplished on these questions. Our results show that such a variable is pertinent, at least to the question of inflection in French. We can therefore consider that some of the conflicting results in the literature might be due to a lack of control of the "lexical environment" of the linguistic material.

4 Conclusion

Our study investigated the role of orthographic neighbourhood size on the recognition of regular French verbs, within the masked priming paradigm. Primes were regular inflections (past and present) of French verbs with different size of orthographic neighbourhood. Our results establish the relevance of an "orthographic neighbourhood size" variable, i.e. the number of orthographically but not morphologically related words a prime can activate, that behave like competitors for the recognition of the target (base form). Further research has to be conducted to determine the role of this variable on derivational and inflectional morphological processing, but the evidence put forward points out to a interesting direction.

References

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language,* **42**, 390-405.

De Jong, N. H., Feldman, L.B., Schreuder, R., Pastizzo, M., & Baayen, (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and language*, **81**, 555-567.

Drews, E. & Zwitserlood, P. (1995). Morphological and orthographic similarity in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance,* **21,** 1098-1116.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10,** 680-698.

Miller, G., & Chomsky, N. (1963). Finitary models of language users. In R. Luce, R. Bush, & E. Galanter (Eds.) *Handbook of mathematical psychology*. New York: Wiley.

Frost, R., Deutch, A., & Forster, K. I. (2000). Decomposing morphologically complex words in a nonlinear morphology. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **26,** 751-765.

Jackendoff, R. (1997). *The architecture of language faculty.* Cambridge, MA: MIT Press.

Giraudo, H., & Grainger, J. (2001). Priming complex words: Evidence for supralexical representation of morphology. *Psychonomic Bulletin and Review*, **8**(1), 127-131.

Grainger, J., Colé, P., & Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language,* **30,** 370-384.

Pastizzo, M. J., & Feldman, L.. B. (2002). Discrepancies between orthographic and unrelated baselines in masked priming undermine a decompositional account of morphological facilitation. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **28,** 244-249.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed model of language acquisition. *Cognition,* **28,** 73-193.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language & Cognitive Processes*, **15**, 445-485.

Marslen-Wilson, W. D., & Tyler, L. K. (1998). Rules, representations and the English past tense. *Trends in Cognitive Sciences*, **2**(11), 428-435.

Moscoso del Prado Martín, F., Deutch, A., Frost, R., Schreuder, De Jong, N. H., & Baayen, R. H. Changing places: a cross-language perspective on frequency and family size on Dutch and Hebrew. Submitted to *Journal of Memory and Language.*

New B., Pallier C., Ferrand L., & Matos R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Psychologique,* **101**, 447-462.

Rumelhart, D. E., & McClelland, J. L. (Eds.). On learning the past tenses of English verbs. In: Parallel distributed Processing, Explorations in the Microstructures of Cognition (Vol. 2): Psychological and Biological Models, MIT Press, 1986.

Segui, J., & Grainger, J. (1990). Priming word recognition with orthographic neighbours: Effects of relative prime-target frequency. *Journal of Experimental Psychology: Human Perception & Performance,* **16,** 65-76.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language,* **37,** 118-139.

Voga, M., & Grainger, J. (2004). Masked Morphological Priming with Varying Levels of Form Overlap: Evidence from Greek Verbs. *Current Psychology Letters: Behaviour, Brain & Cognition,* Vol. 1, No 12, 2004.

Tsapkini, K., Jarema, G., & Kehayia, E. (2002). Regularity revisited: Evidence from lexical access of verbs and nouns in Greek. Brain & Language, **81**, 103-109.