Quantitatively Detecting Semantic Relations

John Nerbonne & Tim Van de Cruys

January 9, 2008

CENTER FOR THE STUDY OF LANGUAGE AND INFORMATION

1

Quantitatively Detecting Semantic Relations: The Case of Aspectual Affinity

1.1 Introduction

The explosion in the creation of text corpora in recent years suggests that the opportunity may be ripe to examine quantitative techniques for their value in semantics.[1] The present paper aims to explore one quantitative technique with an eye toward its potential value in illuminating questions of semantic theory. It is exploratory in nature, and it will not offer definite conclusions or specific advice.

To exploit large text corpora—of a size of the order of magnitude of 10^9 words, we need to employ automatic procedures of analysis. It is unthinkable to work through such volumes of material except by using computer programs. This leads to a difficulty when one's analytical ambitions are semantic, since the semantics of texts is not immediately accessible to automatic procedures.

Perhaps in some glorious future there will be data annotated for semantics to an extent that makes the direct application of quantitative analyses straightforward. But at present we do not have such resources, only fairly large corpora of texts. This means in turn that we need to operationalize our semantic concepts in a way that is amenable to automatic processing. Naturally this has an impact on the sorts of phenomena that can be studied.

In the present study we concentrate on aspect and on the affinities different sorts of adverbials have with some aspectual categories as opposed to others. We operationalize the inherent aspect of verb phrases

[1]We have benefited from discussions with the computational linguistics group at the University of Groningen and from astute comments by two anonymous referees.

by noting the verb heading the phrase, and the class of the durative adverbial by noting the preposition heading the durative adverbial. Both of these steps are subject to some error, but fortunately, quantitative techniques also tolerate noise in characterizations as long as the statistical strength of the association between the category and its operationalization is sufficient, and as long as the "noise" does not systematically favor some analyses over others. We still need to be on guard against using procedures that bias search routines and ultimately, the results of analysis.

1.2 Background

We discuss computational semantics in this section as well as the statistical background needed for detecting aspectual affinities in large corpora.

1.2.1 Computational Semantics

Work in COMPUTATIONAL LINGUISTICS has frequently implemented and applied work from SEMANTIC THEORY, often with special interest for issues concerning disambiguation (Nerbonne, 1996). Since the quantitative turn in computational linguistics, lexical classification and lexical semantics have been the focus of attention in computational work. Schulte im Walde (to appear, 2008/9) reviews a large number of papers which classify verbs using corpus evidence. The focus is on recognizing subcategorization frames, but there has also been work on detecting selectional preferences, semantic roles and diathesis alternations.

It is difficult to collect the statistical information that is useful to lexical semantics, but many quantitative techniques have nonetheless been proposed with an eye toward detecting semantic properties. In addition to the work on semantic roles mentioned above, there has been a substantial number of papers aimed at distinguishing regular semantic combination from the irregular sorts of combination found in multi-word expressions, idioms, and non-compositional constructions (Villavicencio et al., 2005, Deane, 2005, Villada Moirón, 2005).

A second area of focus in quantitative work on semantics has been LEXICAL SEMANTIC SIMILARITY. Automatically acquiring a relative measure of how semantically similar a word is to known words is much easier than determining what the actual meaning is, as Manning and Schütze (2000, § 8.5) point out. Manning and Schütze's textbook refer to a number of works in which the detection of lexical semantic similarity is central.

Brent (1991) shows that one can distinguish stative verbs from others using the presence of progressive variants as well as adverbials expressing rates of speed (e.g., *quickly*), essentially using frequencies of

combination as a cue. Siegel (1998) extends Brent's work, using several more cues, including durative adverbials, which we focus on below. He applied three techniques from statistics and machine learning (logistic regression, decision trees and genetic algorithms) with the goal of classifying the aspectual predication of the clause. Since we attempt to detect the relations between the verbal heads and the durative adverbials, we effectively attempt to classify these simultaneously, and also to gauge the strength of the association between subtypes of these.

Mehler (2007) analyzes semantics from a quantitative perspective and recognizes the need for structural sensitivity in quantitative semantics. He allows for structural effects by viewing semantic combination as a HIERARCHICAL CONSTRAINT SATISFACTION PROBLEM (HCSP). In principle Mehler thus allows that semantic combination be dependent on "syntactic dependency and text coherence relations" (p. 147), even if the specific constraints he handles are discourse based rather than grammatical. Like Mehler, we shall be concerned to rise above the lexical level, but where his focus is on the theoretical underpinnings of the relation between quantitative work and semantic theory, we shall try to develop an experimental technique.

In this paper we try to extend the techniques used in computational linguistics to address questions in semantic theory. Since it is an early effort we will try to be alert for signals that we are detecting semantic structure and to be open for opportunities to exploit the information that large corpora, processed automatically, might offer.

1.2.2 Vector Space Model

The VECTOR SPACE MODEL is one of the most widely used models to investigate lexical semantic similarity in textual data, mainly because it is easy to understand, and it allows one to express 'semantic proximity' between entities in terms of spatial distance (Manning and Schütze, 2000). It is particularly popular in Information Retrieval, where it is used to create term-document matrices. We introduce the vector space model via an example. Consider two documents, one about *Belgium* (B) and one about the *Netherlands* (NL).

- Belgium is a kingdom in the middle of Europe, and **Brussels** is its capital. **Brussels** has a Dutch-speaking and a French-speaking university, but the largest student city is **Leuven**. **Leuven** has 31,000 students.
- The Netherlands is a country in Western Europe, located next to the North Sea. The Netherlands's capital is **Amsterdam**. **Amsterdam** has two universities. **Groningen** is another important

student city. In **Groningen**, there are 37,000 students.

The documents can easily be transformed into a term-document matrix, in which each document is represented by a vector. Each dimension in the vector corresponds to a term (a word or fixed expression), where the frequencies of the terms (in this case, cities) in each of the documents are indicated. The resulting matrix is shown in Figure 1.

$$\begin{bmatrix} & B & NL \\ Groningen & 0 & 2 \\ Leuven & 2 & 0 \\ Amsterdam & 0 & 2 \\ Brussel & 2 & 0 \end{bmatrix}$$

FIGURE 1 A term-document matrix

To term-document matrices like this one, SIMILARITY MEASURES can be applied,[2] to assay how similar documents are to each other, or to a query entered by the user. Note that the *Belgian* document is represented in the term-document matrix as vector of city-reference frequencies, viz. $< 0, 2, 0, 2 >$, the first column in Figure 1. Similarly, words are represented as document-occurrence frequencies, so that *Groningen* is just $< 0, 2 >$. Vector spaces are immediately amenable to the application of distance metrics, which is one reason why they are popular. These distance metrics are used to assay the similarity between words. The two words for which the semantic similarity is to be calculated are represented as vectors in a multi-dimensional feature space.

The co-occurrence information that can be captured by such matrices is not limited to words and documents. If we add grammatical analysis, we can straightforwardly record the dependency relations of a particular word. In that case, the dimensions of the vectors correspond to the dependency relations that the word occurs in together with the lexical head to which it is related. Dependency relations that might be suitable for a word like *apple* could e.g. be 'object of verb *eat*' and 'modified by adjective *red*'. Similarly, such matrices can capture the co-occurrence information that is the subject of this research, viz. verbs and the adpositional heads of modifiers modifying those verbs. Figure 2 gives an example of two adpositions represented as vectors, with some verbs as features. We read '5' in the $(leave, at)$ cell of the matrix, indicating that,

[2]The COSINE is a natural choice when dealing with vectors, but one may also use the inverse of distance measures such as Euclidean distance or MANHATTAN DISTANCE, or set-based measures such as JACCARD.

in our example corpus, *at* occurred five times as the prepositional head of a modifier modifying a verb phrase headed by *leave*.

	leave	*start*	*work*	*live*
at	5	7	0	0
during	0	0	7	6

FIGURE 2 A preposition by verb matrix

The matrix shows that—in our sample corpus—the preposition *at* collocates with the verbs *leave* and *start*, while *during* only collocates with *work* and *live*. A matrix of this kind is the basic input for subsequent statistical techniques (see below). A more detailed explanation about the construction of our co-occurrence matrix can be found in § 1.3.4.

Semantic classification based on co-occurrence frequencies improves significantly when more informative collocations are weighted more heavily. Some features, such as co-occurrences with the verbs *have* and *be*, are less informative because they occur with many words. Other features only occur with a limited number of words, and are thus more informative. To account for these distributional differences, we use POINTWISE MUTUAL INFORMATION (Church and Hanks, 1990). Intuitively, PMI assigns a high value when the frequency of two events co-occurring is much higher than would be expected on the basis of the individual events' frequencies. The formula is given as equation 1.1.

$$I(x, y) = log \frac{p(x, y)}{p(x)p(y)} \tag{1.1}$$

1.2.3 Singular Value Decomposition

SINGULAR VALUE DECOMPOSITION (SVD) is a technique that is used to calculate a so-called *low-rank approximation* of a matrix. It is often used as a dimensionality reduction technique in applications that involve large-scale matrix computations.

Technique

SVD originates from linear algebra: a rectangular matrix is decomposed into three other matrices of specific forms so that the product of these three matrices is equal to the original matrix.

$$A = TSD^T \tag{1.2}$$

where A is the original matrix. The first component matrix, T, contains the same number of rows as the original matrix, but has m columns, corresponding to new, especially derived variables. The second component matrix D has the same number of columns as in the original matrix A, but m rows of derived vectors. These specially derived variables are respectively called left-singular vectors and right-singular vectors. The third matrix S is a diagonal matrix: it is a square $m \times m$ matrix with non-zero entries only along the diagonal. This matrix contains derived constants called SINGULAR VALUES, and these are ordered with respect to their significance in contributing to the product that approximates the original matrix. If one or more of the least significant singular values are omitted, then the reconstructed matrix will be the best possible approximation of the original matrix in the lower dimensional space.

One key property of the derived matrices is that all dimensions are linearly independent: they are *orthogonal* to each other. This is an aid to the interpretation of the results.

Thorough understanding of singular value decomposition requires a firm background in linear algebra. Below we will try to sketch the idea behind SVD intuitively. The interested reader may consult a good introduction on linear algebra (e.g. Strang (2003)) for more information. Landauer and Dumais (1997) also contains a brief but illuminating appendix on SVD.

Example

We will now look at a small, made-up example to see how SVD might be able to detect latent semantic structure present in the data. Figure 3 shows the SVD of the matrix in Figure 1.

$$A \begin{bmatrix} & B & NL \\ Groningen & 0 & 2 \\ Leuven & 2 & 0 \\ Amsterdam & 0 & 2 \\ Brussel & 2 & 0 \end{bmatrix} = T \begin{bmatrix} 0.00 & 0.71 \\ -0.71 & 0.00 \\ 0.00 & 0.71 \\ -0.71 & 0.00 \end{bmatrix} * S \begin{bmatrix} 2.83 & 0 \\ 0 & 2.83 \end{bmatrix} * D^T \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

FIGURE 3 SINGULAR VALUE DECOMPOSITION of a term-document matrix

The original matrix A is decomposed into three other matrices T, S and D^T. The singular values in S show that two equally important

dimensions are found; furthermore, the left- and right-singular vectors show that the frequencies are evenly divided among terms as well as among documents.

Figure 4 shows what happens when we add another document about Belgium, with a slightly different frequency distribution of terms: the Belgian dimension becomes the most important (i.e. captures the most variation, 2.92), while the Dutch dimension remains the same (2.83). The third dimension (0.68) captures the remaining variation (the fact that the third document only talks about Brussels).

$$\begin{bmatrix} & B & NL & B \\ Groningen & 0 & 2 & 0 \\ Leuven & 2 & 0 & 0 \\ Amsterdam & 0 & 2 & 0 \\ Brussel & 2 & 0 & 1 \end{bmatrix} =$$

$$T \begin{bmatrix} 0.00 & -0.71 & \sout{0.00} \\ -0.66 & 0.00 & \sout{0.75} \\ 0.00 & -0.71 & \sout{0.00} \\ -0.75 & 0.00 & \sout{-0.66} \end{bmatrix} * S \begin{bmatrix} 2.92 & 0.00 & \sout{0.00} \\ 0.00 & 2.83 & \sout{0.00} \\ \sout{0.00} & \sout{0.00} & \sout{0.68} \end{bmatrix} *$$

$$D^T \begin{bmatrix} -0.97 & 0.00 & 0.26 \\ 0.00 & -1.00 & 0.00 \\ \sout{-0.26} & \sout{0.00} & \sout{-0.97} \end{bmatrix} \cong A' \begin{bmatrix} 0.00 & 2.00 & 0.00 \\ 1.87 & 0.00 & 0.50 \\ 0.00 & 2.00 & 0.00 \\ 2.12 & 0.00 & 0.56 \end{bmatrix}$$

FIGURE 4 Truncated Singular Value Decomposition

If we now truncate the SVD by keeping only the two most important dimensions, and then reconstruct our original matrix, we get matrix A', which is the best possible reconstruction from only two dimensions. Note that matrix A' resembles matrix A, except for the numbers of the third document: instead of assigning all frequency mass to the term *Brussel*, the mass is almost evenly divided among the Belgian terms *Brussel* and *Leuven*. When keeping only two dimensions, the SVD "guesses" the best possible distribution. This is an example of how the technique is used to obtain more succinct models.

Applications

While rooted in linear algebra, singular value decomposition has proven to be a useful tool in statistical applications. In this respect, it is akin to statistical methods such as factor analysis, correspondence analysis and principal components analysis. The technique can easily be interpreted

statistically: the left-singular and right-singular vector linked to the highest singular value represent the most important dimensions in the data (i.e. the derived dimension—independent of other dimensions—that explains the most variance of the matrix). The singular vectors linked to the second highest value represent the second principal component (orthogonal to the first one), and so on. Typically, one uses only the first n principal components, stripping off the remaining singular values and singular vectors. Intuitively, SVD is able to transform the original matrix—with an abundance of overlapping dimensions—into a new, many times smaller matrix that is able to describe the data in terms of the principal components. Due to this dimension reduction, a more succinct and more general representation of the data is obtained. Redundancy is filtered out, and data sparseness is reduced.

SVD has achieved good results in INFORMATION RETRIEVAL (IR), where it is applied in the framework of LATENT SEMANTIC ANALYSIS (LSA, Landauer and Dumais (1997), Landauer et al. (1998)). In LSA, a singular value decomposition is applied to a fairly large term-document matrix (on the order of 30K terms by 30K documents). According to its proponents LSA finds actual lexical tendencies in the data. The technique is applied in order to obtain a small number of semantic dimensions in terms of which words are characterized for retrieval purposes. This allows researchers (and practitioners) in IR to ignore many other characteristics of individual noun distributions.

The fact that LSA is able to discover some kind of latent structure is shown by its performance on a synonymy test, which is part of a Test of English as a Foreign Language (TOEFL). This is a test given to foreign applicants to American universities. If one applies LSA to determine the closest synonym in a word-pair test in multiple choice format, the algorithm scores as well as the average non-English speaking participant in the test (Landauer et al., 1998).

The calculation of SVD involves iteratively solving a number of eigenvalue problems, which we will not discuss. A number of programs are available that can handle the kind of large-scale singular value decompositions that are necessary for linguistic data sets. In this research, SVDPACK (Berry, 1992) has been used. SVDPACK is a program that is able to handle sparse matrices efficiently and quickly (depending on the number of singular values one wants to retain). Most decompositions can easily be computed on ordinary Unix workstations, in a reasonable amount of time.

1.3 Aspectual Affinity

We turn to the linguistic phenomenon which is the focus of our experiments, viz., the affinity in combination that is seen in English, Dutch and other languages between different verbal aspects (aka aspectual classes or *Aktionsarten*) on the one hand and two different sorts of durative adverbial on the other.

1.3.1 Background

Vendler (1967), building on Aristotelian concepts and twentieth century work by Ryle and Kenny, distinguished four classes of verbs (or verb phrases) based on their logical and grammatical properties: states, activities, achievements and accomplishments. We do not attempt to summarize all of this work, but suggest the sorts of properties that are appealed to. We suggest Dowty (1979), Moens and Steedman (1988) and Egg (2005) for more comprehensive discussion of the differences in linguistic and inferential behavior among the aspectual classes. For example, states such as *be tired* do not normally occur in the progressive **be being tired*, while activities such as *sing* easily can. Accomplishments such as *draw a box* and achievements such as *die* or *notice* are associated with implicit completions or end points, which leads to striking differences compared to activities. For example, someone who stopped drawing a house at a certain time, did not draw one, while someone who stops being tired or stops singing certainly was tired and did sing at an immediately prior time.

States and activities do not have implicit completions or end points and are therefore referred to as ATELIC; both combine with adverbials of duration headed by *for*, unlike the TELIC classes of achievements and accomplishments, which combine with adverbials of duration headed by *in*. While all achievements and accomplishments combine felicitously with *in*-adverbials, if one attempts to combine them with durative adverbials headed by *for*, the resulting phrase is usually understood iteratively. Thus someone who is said to have drawn a box for an hour, is normally understood to have drawn the box repeatedly, not a single time lasting an hour. Compare the combination with *notice* as well.

Some accomplishments combine with adverbials headed by *for* without being understood iteratively, namely those with a clearly associated resulting state. Thus if someone opened a channel for several weeks, we might understand either that it was repeatedly opened (as noted above) or that it was opened once and remained open. The adverbial indicates the length of time it stayed open in that case.

Let's dwell on the parenthetical mention of "verb phrases" in the

first paragraph because its ramifications will influence our chances of finding distributional traces of the distinction. We can illustrate the importance aspect of this extension from verbs to verb phrases with a simple example. *Drink a glass of juice* is a telic predicate with a clearly defined end point, while *drink juice, drink liters of juice* and *drink* used without a direct object, in at least one of its uses, are all atelic. This illustrates why we prefer to speak of telic vs. atelic verb phrases, rather than telic vs. atelic verbs *simpliciter*. If we are to detect signals of telicity among verb phrases in full generality, we need to incorporate information about the object as well (as well as other adverbials the verb phrase is modified by). In fact we will find it necessary below to ignore the presence of objects and other modifiers for reasons we discuss there.

Dowty (1979, Chap. 2) is the *locus classicus* for modern discussion of these topics, where he reviews and discusses the classification as well as the distinctions in logical and grammatical behavior that motivate the classification. He goes on then to suggest underlying differences in temporal reference that help to explain why the distinctions exists, proposing an "aspectual calculus" (§ 2.3) to account for the limited range of combinations and their meanings. Krifka (1987) proposes an axiomatic characterization of the distinction.

1.3.2 Mixed Affinities

We could not hope to do justice to all of the work done in Dowty's *Word Meaning and Montague Grammar* (and following this) on the telic/atelic distinction, but we shall focus on this distinction below as we seek distributional traces of it in a very large corpus. We review a discussion about the strictness of the combinatorial restrictions since this bears directly on the chances of finding distributional traces of the distinction in corpora.

As Dowty (1979, § 2.2.5) notes, one cannot apply the test of adverbial modification to partition verbs (or verb phrases) into distinct classes. It is natural to say that someone *read an article in an hour*, but not ill-formed to say that someone *read an article for an hour*. Thus it is possible to construe naturally telic phrases such as *read a book* as atelic in some circumstances. Similarly, normally atelic verbs such as *swim* can also be construed telically. Dowty (1979, p.61) notes that, if interlocutors know that John swims a certain distance daily, it can be natural to say that *John swam today in an hour*, for example, to indicate how long it took him to swim his normal distance. Dowty (1979) notes that the other linguistic tests indicating telicity likewise apply, but we do not repeat these here.

These sorts of considerations led Moens and Steedman (1987) to construe aspectual classes as types and to investigate the role of various temporal operators as type-changing. The progressive aspect, seen from this perspective, maps a telic accomplishment such as *write a book* to an atelic activity capable of combining with an atelic-seeking durative such as *for a year* to obtain the felicitous *was writing a book for a year*, which indeed does not imply that the book was completed.

Moens and Steedman (1987) note that applying an adverbial of an inappropriate type such as *in an hour* to a verb such as *swim* might be viewed as TYPE COERCION, an analytical possibility examined at length by Egg (2005), who rejects type coercion as a general account, appealing instead to LANDING SITE COERCION to explain the (most accessible) reading of *She left for an hour* as meaning 'she left and stayed away for an hour'.

1.3.3 Distributional Expectations

The basic generalization is very simple, namely that atelic verbs (and verb phrases) combine with durative adverbials headed by *for* and telics with adverbials headed by *in*, but as we have seen there are a number of necessary qualifications. First, for many verbs the nature of the grammatical direct object influences whether it refers to a telic or atelic event. Second, our knowledge of the expected length of an event may likewise influence our judgment as to whether the event should be understood telically or atelically. For example, *read a book* refers to a telic event, but if one hears that *Sue read 'War and Peace' this morning*, one is inclined to understand it atelically, given what one knows of the book's length, and limits on reading speeds. Third, some adverbials, e.g., *this morning*, can combine with both telic and atelic verb phrases. The research literature does not suggest distributional restrictions that are hard and fast.

On the other hand, the basic generalization sounds at first quite convincing, and the exceptions often sound a bit strange. Adverbials headed by the preposition *in* or, alternatively, by the preposition *for*, combine to form telic and atelic verb phrases respectively, sometimes "coercing" the verb phrases they combine with into the right aspect. This suggests that the relative affinities of the two types of adverbial for the different verbs ought to be reflected in the relative frequency with which combinations are founded, and that the frequency of combinations respecting the affinity ought to outpace that of exceptions. We have additional reasons for ignoring the presence of objects (of particular sorts that influence aspect) in verb phrases, namely first that it is difficult to recognize the relevant range of objects automatically, and second that including verb-

object combinations would increase the number of categories greatly, reducing the frequencies with which classes are instantiated and thereby the reliability of the analysis.

Furthermore, since our primary aim in this paper is to explore the possibility of applying quantitative techniques in order to detect semantic structures, we wish to forge ahead in spite of the potential exceptions. We wish to verify whether such an affinity is reflected in the very large corpora we examine; whether we can detect aspectual classes among the verbs heading verb phrases in construction with specific durative adverbials; whether we can detect classes of temporal adverbials with aspectual affinities like those of adverbials headed by *in* and *for*, respectively (and what those prepositions in fact are); and, finally, whether there are other conditioning factors on telicity. It will be interesting if it turns out that some adverbials prefer to combine with the one or the other aspectual class, even though there is no straightforward aspectual requirement that they do so. This putative phenomenon has not received theoretical attention, but the quantitative approach we apply lends itself naturally to the question.

1.3.4 Methodology

In order to explore semantic distinctions among verbs and temporal adverbials in a quantitative way, the required frequency information needs to be extracted from a corpus. First of all, a set of 22 adpositions that occur as head of temporal adverbial PPs was manually selected. These adpositions and their translations are shown in Table 1.

adposition	translation	adposition	translation
in	'in'	*geleden*	'ago'
na	'after'	*later*	'later'
over	'in'	*lang*	'for'
binnen	'within'	*op*	'in'
sinds	'since'	*voor*	'for'
om	'at'	*tot*	'until'
eerder	'before'	*vanaf*	'from'
gedurende	'during'	*terug*	'ago'
door	'during'	*rond*	'around'
halverwege	'half way'	*tijdens*	'during'
tegen	'around'	*achtereen*	'in a row'

TABLE 1 Manually selected adpositions that head time adverbials

Many of these adpositions also appear in PPs that are not temporal

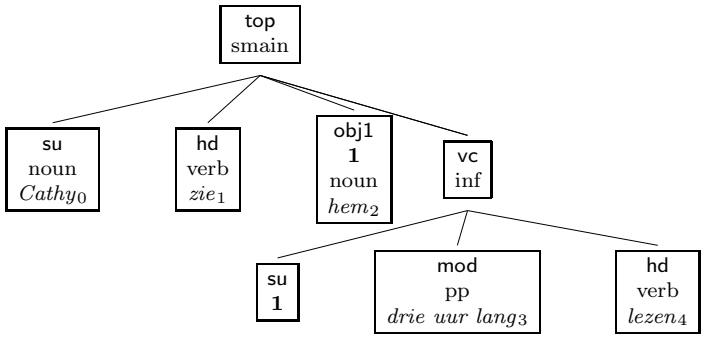

FIGURE 5 Example of dependency structure for the sentence *Cathy zag hem drie uur lang lezen* 'Cathy saw him read for three hours'

adverbials. To make sure that only temporal adverbials are extracted, we only take into account PPs that contain an NP headed by nouns that express a time quantity, namely *minuut* 'minute', *uur* 'hour', *dag*, day', *week* 'week', *maand* 'month' or *jaar* 'year'.

The combinations of verbs and modifying time adverbials were automatically extracted from the *Twente Nieuws Corpus*,[3] a large corpus of Dutch newspaper texts (500 million words), which has been automatically parsed by the Dutch dependency parser Alpino (van Noord, 2006). Alpino reaches an accuracy of up to 90% per grammatical dependency. An example parse is shown in Figure 5. Alpino's output parses are saved in XML. An XSL-style-sheet was developed to extract the required information, viz. the adposition and the verb (in this example the postposition *lang* and the verb *lezen*).

Next, a matrix is created of the occurrences of the 22 temporal adpositions cross-classified by the 5,000 most frequent verbs. The matrix contains the frequency of each co-occurrence. Each value is weighted with pointwise mutual information (see above), so that more informative features get a higher weight.

As a last step, singular value decomposition is applied, reducing the matrix to two dimensions. This dimensionality reduction exposes the two most important dimensions that are present in the data. Additionally, using two dimensions has the advantage that the results can easily be visualized. We should note that other dimensions can, in principle, be interesting as long as they contribute significantly to the explanation of variance. But we found it difficult to interpret the third and following dimensions, whose importance naturally decreases as well.

[3] `http://www.vf.utwente.nl/~druid/TwNC/TwNC-main.html`

Since the parsing method is fully automatic, our input may contain erroneous parses. However, our statistical techniques are robust with regard to random noise. Only systematic noise (introduced by the parser's grammar or disambiguation model) can cause problems in the analysis. A random evaluation of the parsing results turned up no indication of systematic error.[4]

1.3.5 Results

From the inspection of the results, we conjecture that the first SVD dimension gives an indication of the (a)telicity of the clause, and the second dimension of the duration of the timespan (introduced by the durative adverbial). We will concentrate on the first dimension in examining the data.

Adverbials

Figure 6 gives a graphical representation of the adverbials in the reduced dimensionality space; the x-axis shows the first dimension, and the y-axis shows the second dimension. The first dimension captures approximately 9.5% of the variance present in the original matrix, the two first dimensions together account for approx. 15.5%.

Due to the application of the SVD, a continuum emerges between atelic modifiers (*lang* 'for') and telic modifiers (*om* 'at [3 o'clock]', *rond, tegen* 'around [3 o'clock]', and *geleden* '[several hours] ago'). We note that *binnen* 'within [an hour]' is also correctly classified toward the telic end of the spectrum, as is *over* 'in [a week]'. But we also note three examples which do not seem immediately plausible, given our interpretation of the first dimension as (a)telicity. We shall examine *achtereen* '[three days] in a row', *tijdens* 'during', and *geleden* 'ago'. *Achtereen* seems as if it should be used iteratively and therefore atelically just by virtue of its meaning.[5] *Tijdens* can be crucially used both with an event specification, e.g. *tijdens de oorlog* 'during the war', and also with a specification of temporal durations, e.g. *tijdens drie dagen* 'for a period of three days'. The latter should be atelic, and the former might be compatible with either telic or atelic propositions. Both of the prepositions are found fairly far to the left on the x-axis, among predominantly telic prepositions.

[4] As parsing experts are well aware, the attachment of prepositional phrases is subject to more error than other parse decisions. But we know of no tendency for parsers to err more in attaching telic durative adverbials than in attaching atelic duratives, for example.

[5] In fact *achtereen* can also be used with no hint of iterativity. *Hij werkte hier drie jaar achtereen* 'He worked here for three years in a row'. But we suspect that iterative uses predominate.

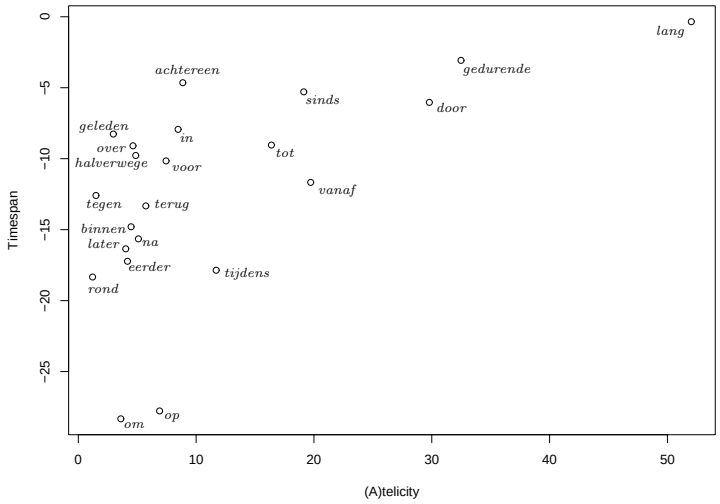

FIGURE 6 The classification of the heads of temporal adverbials according to the two most significant dimensions of the SVD applied to the frequency co-occurrence matrix between such heads of adverbials and heads of verb phrases in 400 million words of Dutch newspaper text.

Finally we shall examine *geleden* 'ago', which combines with expressions of duration such as *drie jaar* 'three year(s)' to form frame adverbials. Since there is no logical constraint restricting the combinations of frame adverbials with either telic or atelic verb phrases, we expect frame adverbials to be relatively insensitive to aspect—both *lived in Columbus a decade ago* but also *ran a mile a week ago* are well-formed. It is therefore surprising to see *geleden* classified as having such a strong affinity with telicity (see Fig. 6). Similarly *op* heads frame adverbials once combined with an expression denoting a day: *op zondag* 'on Sunday', or *op Pasen* 'at Easter'. It should likewise be expected to be somewhat neutral with respect to telicity, but it too is classified much more closely to the telic extreme.

To test whether these interpretations are plausible, we decided to examine a random sample of twenty sentences for each of the prepositional or postpositional heads from the material that the SVD is based on. We shall not present all twenty sentences, but only the verbs and arguments that are crucial for the category telic/atelic.

We first examine twenty random sentences involving adverbials headed

verb phrase	argument
suffered pain in her abdomen	during an au-pair year
168 rebels were killed	during the first six days
aim at a number of subscribers	during the first year
warm air was introduced	during the first week
compose music	during his Paris years
contract a cold	during the last few days
drop marines off	during the first day of [...]
what goes on	during a moment of silence
fill the auditorium	during the first three days
grow	[during] the most recent days
spend time	during his last [work]days
play great tennis	during the last two weeks
remain the same	during your 12 years as [...]
not function well	during the last years
record stories	during the days in [...]
[this] wine was bottled	during the last year of [...]
promote sport	during their days
work with [...]	during this hectic month
get lost	during the longest and hardest day
maintenance costs are low	during the first two years

TABLE 2 Crucial verb-argument combinations in randomly sampled sentences (from twenty million) involving aspectual adverbials headed by *tijdens* 'during'. Most of the predications ought to be classified as atelic, with, however, exceptions such as 'contract a cold', 'be bottled (of (a specific quantity of) wine)', and 'get lost'.

by the preposition *tijdens* 'during'. The adverbials, the verb and the relevant arguments are listed in Table 2. The table (as well as the full examples, not shown) seem to bear out the linguistic intuition that adverbials formed with *tijdens* combine primarily with atelic predications. This means in turn that, if our hypothesis about this first dimension is correct, i.e., if the x-axis in Fig. 6 does correspond with telicity, then the boundary for adverbial heads that combine with telics lies to the left of this point.

We turn now to the examination of twenty random sentences involving adverbials headed by the postposition *achtereen* '[three days] in a row'. We note that only one of the twenty sentences was crucially misparsed. In that case a temporal adverbial that modified a noun phrase was parsed as modifying the verb. We found the combinations shown in

Table 3.

The range of examples in Table 3 illustrate how difficult it is to deal with input that has not been selected for its clarity with respect to theoretical issues. 'Fill the auditorium' seemed to be used atelically in the sentence in the example. Several examples involve clearly telic predications which, probably because of appearing in construction with the postposition *achtereen*, are nonetheless understood atelically. This suggests that SVD is classifying the postposition as telic due to the fact that most of the material it combines with is at base telic. Other examples are more complicated, involving potential telics in construction with quantified arguments, leading to an atelic reading (e.g., 'repel everything', 'cost sales', or 'watch ads'). Interestingly enough, there are two examples where fundamentally atelic predications are made telic via subordinate modification ('jog a quarter hour' and 'wear a jersey for several days'). But there are also examples of atelic predications with no relevant type-shifting operators ('enjoy a high salary', 'sit in the car' and 'read'). But perhaps we impute too much structure in discussing the possibility of type shifting here. The most correct conclusion may be just that adverbials with iterative meaning apply felicitously to both telic and atelic predications.

Third, and finally, we examine twenty random sentences involving adverbials headed by *geleden* 'ago'. In order to save space, we do not list the predications and adverbials fully, but the report is quite simple: virtually all (19) of the twenty examples involve telic predications! This is interesting for two reasons. First, it is consistent with our hypothesis, advanced in the discussion of *tijdens*, that the border of telic vs. atelic may lie fairly far to the left in the scatterplot—in fact it ought to lie between these two adverbial heads.

But it is also interesting because it suggests that the affinities between adverbials and aspect are not exhaustively described by aspectual theory. As we noted above, there is no reason why a postposition such as *geleden*, which combines with an expression denoting a length of time to form a frame adverbial, should show a preference for one aspect over another. In particular, this sort of tendency seems to have nothing to do with the principles of aspectual interpretation based on type coercion that Moens and Steedman (1987) examined, nor with the sorts that Egg (2005) calls LANDING SITE COERCION. Since, however, predications involving *geleden* are always about the past, we perhaps see more telic predications because there is a tendency to view past actions as completed.[6]

[6]Comrie (1976) discusses interrelations between aspect and tense.

verb phrase	argument
production fell	three years in a row
won the championship	three years in a row
'enjoy' a high salary	for many consecutive years
repel everything [every attempted goal]	for 908 consecutive minutes
jog a quarter hour	for three days in a row
show low growth	for months on end
cast his fishing rod	for many years in a row
wear the yellow jersey for several days	three years in a row
be clear to me	four days in a row
presents [herself] as hotel slut	five days in a row
cost the company sales	five years in a row
participate in the championships	four years in a row
watch TV ads	for four hours in a row
sing it	three years in a row
fall 2,75% in purchasing power	four years in a row
sit in the car	for 48 hours in a row
read	for several hours in a row

TABLE 3 Crucial verb-argument combinations in nineteen randomly sampled sentences (from twenty million) involving aspectual adverbials headed by *achtereen*. It is striking that many are telic predications used iteratively, and therefore atelically, while others are atelic by virtue of plural arguments ('watch ads'). At the same time, it is clear that some examples are simply atelic and misclassified ('read').

Verbs

Despite some irregularities noted in the previous paragraphs, our data analysis of the adpositions shows a clear telicity continuum emerging from left to right. The next step is to plot the verbs against the same dimensions, to see whether the same continuum emerges, and if so, where the verbs appear in this continuum. In Figure 7, a sample of the 100 most frequent verbs is plotted in the first two dimensions.

Figure 7 shows the same tendency as Figure 6: a telicity continuum emerges from left to right, with typically telic verbs (*meld* 'report', *vertrek* 'leave' and *begin* 'begin') on the left, and typically atelic verbs (*werk* 'work', *blijf* 'stay' and *volg* 'follow') on the right.

To investigate the continuum for verbs, we again examined 20 randomly selected sentences for a number of verbs. Table 4 shows a number of sentences with the verbs *werk*, *blijf* and *volg*, appearing at the atelic end of the continuum. Practically all of the sentences indeed have an

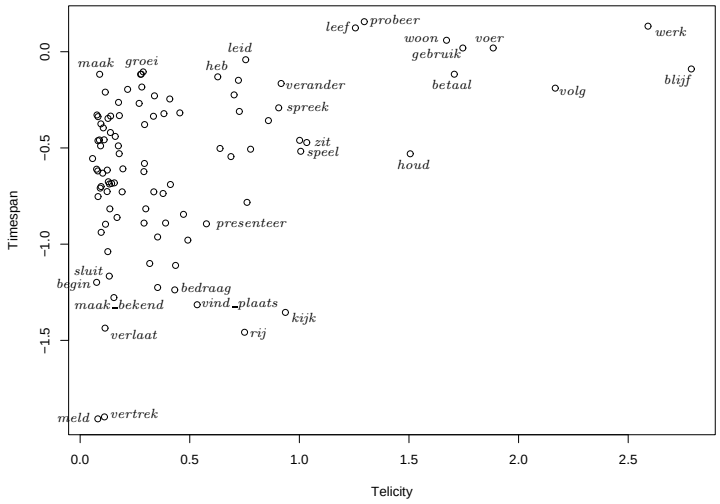

FIGURE 7 A graphical representation of verbs, showing the first two dimensions of the SVD

atelic reading; a few telic interpretations are also present, such as 'a training [session] followed around 6 o'clock'. It is revealing that these telic instances are only present with the verb *volg*, which appears already more to the left in the continuum.

verb phrase	argument
all immigrants [...] have worked	for three to five years
Rotterdam works with health care information	since one year [ago]
He wanted to work half time	after a year
It stays dry	for five days in a row
Bob Dylan stayed my hero	throughout the years
The reporters followed the three women	during one year
... followed by the FBI	since one year [ago]
a training followed	around 6 o'clock

TABLE 4 Some verb-argument combinations for the verbs *werk*, *blijf* and *volg*, appearing at the atelic end of the continuum.

The other end of the spectrum shows the opposite picture: the majority of the sentences selected for the verbs *meld* 'report' and *vertrek*

'leave' are telic; atelic readings are possible with atypical cases, as in 'he reported ... for 20 minutes' (Table 5).

verb phrase	argument
Lindsay reported [herself]	a week and a half ago
124 voters reported	at 2 o'clock in the afternoon
The Greek media reported	on the day of her funeral
The man left	at 9 o'clock in the morning
Fricke left again	half an hour later
The hostage started	before 8 o'clock
He reported ...	for 20 minutes

TABLE 5 Some verb-argument combinations for the verbs *meld*, *vertrek* and *begin*, appearing at the telic end of the continuum.

Next, we have evaluated two verbs that are located towards the middle of the continuum, namely the verbs *speel* 'play' and *kijk* 'look'. Some crucial sentences are shown in Tables 6 and 7. Both tables show a mix of telic and atelic verb phrases, in line with the telicity continuum. The first two sentences in Table 6 are telic instances; the five other sentences are atelic. In Table 7, the three first sentences are telic, the remaining four atelic.

verb phrase	argument
The band started playing	shortly after 8 o'clock
... the second game is played	within 14 days
He played in Greece	during those 3 years
No senior competitions are played anymore	since last year
... where the Yankees were playing	three days in a row
... who played with 9 players	for an hour
This station has played an important role	throughout history

TABLE 6 Some verb-argument combinations for the verb *speel*

1.4 Conclusion & Future work

This paper took the availability of increasingly large corpora as an impetus to investigate a statistical technique that seems suitable for investigating "latent" semantic phenomena quantitatively without requiring a manually annotated corpus. The aspectual phenomenon is latent in that it is not reflected simply in overt text; it is reflected indirectly in the

verb phrase	argument
Sluis himself went to look	two years ago
100.000 came to watch the movie	during the first 4 days
We looked at the same place	an hour later
... to watch the Dutch team	during the night
[We'll be] watching video's again	for one month
We looked what the possibilities are	during the last two years
... to watch tv commercials	4 hours in a row

TABLE 7 Some verb-argument combinations for the verb *kijk*

sets of verbs and adpositions used. Our chosen technique relies on the availability of massive amounts of automatically annotated data, but it does not require that the annotations be flawless.

We have analyzed the co-occurrence data of verbs and aspectually sensitive adverbials in an attempt to verify that the well-studied affinity of atelic vs. telic predications for combination with only certain duratives could be detected, and, naturally, the degree to which it held. The data analyzed consisted of parsed sentences from which the verb heading the verb phrase was extracted, and also the preposition or postposition heading the durative adverbial. We work from large tallies showing which verbs were found in construction with which temporal adpositions.

We applied singular value decomposition to extract simultaneously the most significant principal components of the adpositions and also of the verbs. The first two correspond with atelicity on the one hand, and length of the time duration on the other. By examining random samples of data in which different adverbials and different verbs occur, we were able to confirm this interpretation, and also to note that the characterizations are not categorical, but gradual—we found aspectual mismatches in all of the sets of sentences we examined. From semantic theory we know that mismatch is possible, but this study suggests that it also occurs frequently. Interestingly, one of the cleanest sets of examples did not involve aspectually selective adverbials, but rather the aspectually aselective adverbials formed with the postposition *geleden* 'ago', which co-occur almost only with telic phrases. The SVD analysis brought this out.

We conclude therefore that SVD allows us to to generalize over the noisy data acquired from automatic parsing, and it successfully exposed the aspectual tendencies in the telic and atelic verb phrases as well as in the adverbials they construct with.

In future work, we plan to examine alternative methods of statistical analysis. Particularly, methods that bring about an entropy-based dimension reduction (such as NON-NEGATIVE MATRIX FACTORIZATION (Lee and Seung, 2000)) look quite promising. In a similar vein, we want to experiment with different weighting schemes (apart from PMI).

Further applications of this sort of analysis might also involve categories that are not simply marked in data, and for which the issue of semantic conditioning arises, e.g. negative or positive polarity items, or scope relations that are marked by these sorts of items. Further steps in these analyses might involve the examination of alternative hypotheses about the occurrence and co-occurrence of aspectual adverbials, and, indeed, the development of a more rigorous mode in which hypotheses might be tested using these techniques.

Bibliography

Berry, Michael. 1992. Large scale singular value computations. International Journal of Supercomputer Applications 6(3):13–49.

Brent, Michael. 1991. Automatic semantic classification of verbs from their syntactic contexts: An implented classifier for stativity. In Proceedings of the 5th Meeting of the European ACL, pages 222–226. Shroudsburg, PA: ACL.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. Computational Linguistics 16(1):22–29.

Comrie, Bernard. 1976. Aspect. Cambridge: Cambridge University Press.

Deane, Paul. 2005. A nonparametric method for extraction of candidate phrasal terms. In Proceedings of the 43rd Annual Meeting of Association for Computational Linguistics, pages 605–613. Shroudsburg, PA: ACL.

Dowty, David. 1979. Word Meaning and Montague Grammar. Dordrecht: Reidel.

Egg, Markus. 2005. Flexible Semantics for Reinterpretation Phenomena. Stanford: CSLI.

Krifka, Manfred. 1987. Nominal reference and temporal constitution: Toward a semantics of quantity. In Proceedings of the 6th Amsterdam Colloquium, pages 153–173. Amsterdam: Institute of Language, Logic and Information.

Landauer, Thomas and Se Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychology Review 104:211–240.

Landauer, Thomas, Peter Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. Discourse Processes 25:295–284.

Lee, Daniel D. and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In NIPS, pages 556–562.

Manning, Christopher and Hinrich Schütze. 2000. Foundations of Statistical Natural Language Processing. Cambridge, Massachussets: MIT Press.

Mehler, Alexander. 2007. Compositionality in quantitative semantics. In A. Mehler and R. Khler, eds., Aspects of Automatic Text Analysis, vol. 209 of Studies in Fuzziness and Soft Computing, pages 139–170. Berlin: Springer.

Moens, Marc and Mark Steedman. 1987. Temporal ontology in natural language. In 25th Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference, pages 1–7. Morristown, NJ: Association for Computational Linguistics.

Moens, Marc and Mark Steedman. 1988. Temporal ontology and temporal reference. Computational Linguistics 14(1):3–14.

Nerbonne, John. 1996. Computational semantics—linguistics and processing. In S. Lappin, ed., Handbook of Contemporary Semantic Theory, pages 459–82. London: Blackwell Publishers.

Schulte im Walde, Sabine. to appear, 2008/9. The induction of verb frames and verb classes from corpora. In A. Lüdeling and M. Kytö, eds., Corpus Linguistics: An International Handbook, page Chap. 61. Berlin: Mouton de Gruyter.

Siegel, Eric V. 1998. Linguistic Indicators for Language Understanding: Using Machine Learning to Combine Corpus-Based Indicators for Aspectual Classification of Clauses. Ph.D. thesis, Columbia.

Strang, Gilbert. 2003. Introduction to Linear Algebra. Wellesley: Wellesley-Cambridge Press.

van Noord, Gertjan. 2006. At Last Parsing Is Now Operational. In P. Mertens, C. Fairon, A. Dister, and P. Watrin, eds., TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles, pages 20–42. Leuven.

Vendler, Zeno. 1967. Linguistics in Philosophy. Ithaca: Cornell University Press.

Villada Moirón, Begoña. 2005. Data-Driven Identification of Fixed Expressions and their Modifiability. Ph.D. thesis, University of Groningen.

Villavicencio, Aline, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. Computer Speech and Language 19(4):365–377.