

Log-linear models and latent semantic indexing applied to MWE identification.

Begoña Villada Moirón
Alfa-Informatica
University of Groningen
villada@let.rug.nl

Thursday 8th, December 2005

Abstract

A short introduction characterizes the task of identification of multiword expressions and their idiosyncratic properties. Then, this document gives a detailed description of loglinear models and latent semantic analysis. The description enumerates components of the models, estimation techniques for the model parameters and addresses the interpretation of the models and their evaluation. We also briefly report on how the models have been earlier used in identification of multiword expressions. Furthermore, we describe how these models can be used in R statistics package.

A small case study reports on preliminary experiments on quantifying two linguistic attributes. We explore to what extent these attributes identify MWEs and whether a statistical dependence between them can be captured by a loglinear model.

Contents

1	Introduction	2
1.1	What are multiword expressions?	2
1.2	Towards a successful identification approach	3
2	Statistical modeling	3
3	Loglinear models	4
3.1	Preliminaries	4
3.2	Model for a two-dimensional contingency table	6
3.3	Larger two-dimensional tables	8
3.4	Model for three-dimensional tables	10
3.5	Parameter estimation	11
3.6	Evaluation procedure	12
3.7	Loglinear models and MWE identification	12

4	Latent semantic analysis	13
4.1	Classifying multiword expressions using LSA	13
5	Software to apply these models	14
6	Case study: Quantifying linguistic properties	14
6.1	MWE candidates extraction	15
6.2	Quantifying linguistic properties	15
6.3	Are the quantified properties discriminating?	18
6.4	Measuring dependencies with a loglinear model	20
6.5	Summary	21

1 Introduction

In modeling a subset of Dutch multiword expressions (MWEs) – support verb constructions – Villada Moirón (2005) showed that corpus patterns with a relatively high lexical affinity among the component words are not necessarily true support verb constructions. Common association measures such as log-likelihood and salience wrongly rank frequent combinations so called ‘anticollocations’ as good SVCs. The accuracy of an identification model can be increased by taking into account other properties of the expressions such as syntactic and semantic irregularities. In fact, it was shown that knowledge of the syntactic contexts in which SVCs do not (typically) occur helps to reduce the errors made by an automatic identification method. Contexts such as PP over verb, scrambling, nominalization pattern, coordination and pronominalization are discriminating attributes of support verb constructions and other idiomatic expressions.

Here, our aim is to find a statistical technique that incorporates information about discriminative contexts. Such contexts need to distinguish the linguistic behavior of MWEs from the linguistic behavior of regular phrases.

1.1 What are multiword expressions?

Multiword expressions (MWEs) are word combinations whose behavior cannot be inferred from the inherent properties of their component words; these expressions exhibit certain idiosyncrasies at the lexical, morphological, syntactic and/or semantic level.

At the lexeme level, individual lexemes mutually select each other with little lexical variation or room for replacement. A strong lexical affinity holds between the component words in the MWE.

In some lexicalized expressions, productive morphology rules fail to apply to the component words. Nouns in MWEs often show defective morphology (though exceptions exist).

In syntax, a lack of flexibility is commonly observed in MWES. However, some MWES allow various syntactic processes such as determiner alternation, insertion of modification, passivization, etc.

Concerning semantics, the meaning of a MWE is not fully compositional. The meaning of these expressions ranges from fully opaque and noncompositional meaning to almost transparent and compositional meaning.

These are the most common characteristics that have been observed in MWES.

1.2 Towards a successful identification approach

Even though we aim at modeling the phenomenon of MWES, in our study we take a sample from a corpus at a stage when we ignore which expressions are MWES and which ones are regular phrases. The task at stake is to design a method that distinguishes actual MWES from regular phrases automatically; More precisely, we aim at a method that finds all expressions e in the set of linguistic expressions (L) existing in a corpus such that e exhibits the characteristic behavior of MWES. Formally, we seek to delimit the subspace M with $M \in L$.

In order to identify MWES, a probabilistic model should capture the characteristic linguistic behavior of MWES. Thus we aim at finding a technique that helps us establish a set of discriminative linguistic features of MWES not observed in regular phrases and also, dependencies between them.

Our approach goes as follows:

- empirically identify a group of attributes $a \in A$ observed in the elements of M that are less frequent or absent in the elements of L . Each instance of an element $l_i \in L$ is described as a set of k attribute-value pairs a_i of the form $\langle (P, N, V), a_1, a_2, \dots, a_k \rangle$.
- quantify each attribute
- incorporate attributes into a loglinear model
- study statistical association between attributes (these are interaction parameter values). Establish which dependencies are observed in M elements and not (or to a lesser extent) in $L - M$ elements. Are the dependencies more significant in *expressions* $\in M$ than in *expressions* $\in L - M$? If this is the case, we have identified a set of discriminative attributes of MWES.

2 Statistical modeling

Statistical modeling is a way of using quantitative information to make reasonable guesses about the possible relationships among different phenomena that a researcher investigates.

Statistical models differ from other theoretical models in that they recognize a component of chance as an integral part of the model. The components of statistical models are typically quantified: the size or strength of a relationship between the objects of the researcher's interests are represented by numbers, and the numbers can be entered into a formula (which itself is actually the statistical model proper) in order to make predictions about what might happen in a given set of circumstances.

Statistical models are evaluated against a benchmark corresponding to the notion of chance; by exploiting the theory of probability, it is possible to make decisions about which of a set of models are better than the others, and whether in fact any statistical model other than pure chance is needed at all.

Generalized linear models (GLMs) are commonly used to analyse interactions between explanatory and response variables. GLMs include logistic regression and ANOVA typically applied for continuous response variables and, other models for categorical response variables. Logistic regression models are used for binary data i.e. the categories of a variable are best described by a binomial distribution showing only two values (yes/no,1/0). Loglinear models are used for Poisson data or count data.

Logistic regression and loglinear models have been used as a tool for constructing and evaluating quantitative models of relationships in linguistic data. A summary of applications of these models in linguistics can be found in (Paolillo 2002).

3 Loglinear models

One straightforward method for analyzing data is via cross-tabulation. The observations (object of investigation) can be summarized in a multi-way frequency table, usually a table with as many factors as the number of variables taken into account in the research. The loglinear model can be useful to examine cross-tabulation tables, in particular, when modeling relationships between two or more categorical variables (Agresti 2002). This is actually our main reason to choose loglinear models as a probabilistic model. A second reason is that when the number of variables is not too large, interpretation of the interaction effects and model parameters is rather intuitive.

The following description of loglinear models for tables of different dimensions is based on the treatment of these models by Bishop, Fienberg and Holland (1984).

3.1 Preliminaries

To specify the cell probabilities of a two-dimensional 2x2 table such as Table 1, we may use the following constraints:

A_1	A_2		marginals
	1	2	
1	p_{11}	p_{12}	P_{1+}
2	p_{21}	p_{22}	P_{2+}
Total marginals	P_{+1}	P_{+2}	1

Table 1: This 2X2 contingency table represents the cell probabilities of a random sample of cases classified according to two variables A_1 and A_2 .

1. the cell probabilities sum to 1; this is a linear constraint of the form $\sum_{i=1}^2 \sum_{j=1}^2 P_{ij} = 1$;
2. knowing one of the row marginal probabilities and one of the column marginal probabilities: P_{1+} and P_{+1} , we can define all marginal probabilities of the table;
3. one more constraint involving the internal cells is needed to completely specify the structural relationships in the table. Most questions of interest are concerned with the difference between such internal probabilities and the marginal probabilities. Various functions are commonly used (difference of proportions, diagonal sum, ratio of an elementary cell, etc.) but one has desirable properties that the others lack: the cross-product ratio α :

$$\alpha = \frac{P_{11}P_{22}}{P_{12}P_{21}}$$

The cross-product ratio or 'odds ratio' attains the value of 1 when the condition of independence holds; it is invariant under the interchange of rows and columns and, it also remains invariant under the row and column multiplications (the α value remains the same). The odds ratio is an important function in statistical modelling. In certain models, parameter calculation is based on the odds ratio.

Interpretation of the odds ratio The logarithm of the relative odds is also a linear contrast of the log-probabilities of the elementary cells:

$$\log \alpha = \log P_{11} - \log P_{12} - \log P_{21} + \log P_{22} \quad (1)$$

In addition, when $\log \alpha = 0$ we have independence between variables. The above formulation of the $\log \alpha$ suggests that in order to completely specify all cells of the table we look for a linear model in the logarithmic scale.

An example Table 2 shows the cross-classification of a VERB with a PREPOSITION NOUN phrase. The table summarizes the frequency distribution of the verb *stellen* and the phrase *aan de orde*. Applying the formula given above (1), we compute the $\log \alpha$ of the expected counts in order to find out whether a statistical association holds between the two variables. The $\log \alpha(\text{aan_orde}, \text{stellen})$ is 2.914 suggesting that a dependence exists between the variables. Other tables whose $\log \alpha$ is close to zero are observed for the bigrams (*ben, in-jaar*) (0.073), (*maak, in-plaats*) (0.06), (*geef, op-moment*) (0.022), etc. These are expressions where the PP in question does not necessarily co-occur with one (or a few) particular verb(s). The low $\log \alpha$ value assigned to these combinations indeed suggests an independence between the bigram component words.

verb	Preposition+Noun		marginals
	aan_orde	\neg aan_orde	
stellen	2142	12794	14936
\neg stellen	1076	397575	398651
Total marginals	3218	410369	413587

Table 2: Cell expected counts result of cross-classifying two variables 'verb' and a 'preposition+noun'.

3.2 Model for a two-dimensional contingency table

By analogy with analysis of variance (ANOVA) models, a simple way to construct a linear model in the natural logarithms of the cell probabilities is

$$\log P_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)} \quad i = 1, 2; j = 1, 2 \quad (2)$$

where μ is the grand mean, $\mu_{1(i)}$ and $\mu_{2(j)}$ are row and column effects and $\mu_{12(ij)}$ an interaction effect between the two variables:

$$\mu = \frac{1}{4}(\log P_{11} + \log P_{12} + \log P_{21} + \log P_{22}) \quad (3)$$

$$\mu + \mu_{1(i)} = \frac{1}{2}(\log P_{i1} + \log P_{i2}) \quad (4)$$

$$\mu + \mu_{2(j)} = \frac{1}{2}(\log P_{1j} + \log P_{2j}) \quad (5)$$

Since $\mu + \mu_{1(i)}$ and $\mu + \mu_{2(j)}$ are deviations from the grand mean μ , (6) is true. Similarly, $\mu_{12(ij)}$ (interaction term), represents a deviation from $\mu + \mu_{1(i)} + \mu_{2(j)}$, so that, the constraint (7) holds true. The additive properties imply that each μ term has one absolute value for dichotomous variables.

$$\mu_{1(1)} + \mu_{1(2)} = \mu_{2(1)} + \mu_{2(2)} = 0 \quad (6)$$

$$\mu_{12(11)} = -\mu_{12(12)} = -\mu_{12(21)} = \mu_{12(22)} \quad (7)$$

Computing the main effects If $l_{ij} = \log P_{ij}$, the grand mean is the average value of a cell's log probability (8). The row main effects (9) are computed as the deviation of the row marginal log probabilities from the grand mean. The column main effects (10) are computed as the deviation of the column marginal log probabilities from the grand mean. (With the main effects we know which row and column deviates more or less from the grand mean.) The interaction term $\mu_{12(ij)}$ (11) captures any statistical dependence between variables 1 and 2.

$$\mu = \frac{l_{++}}{4} = \sum_{i,j} \frac{l_{ij}}{4} \quad (8)$$

$$\mu_{1(i)} = \frac{l_{i+}}{2} - \frac{l_{++}}{4} \quad (9)$$

$$\mu_{2(j)} = \frac{l_{+j}}{2} - \frac{l_{++}}{4} \quad (10)$$

$$\mu_{12(ij)} = l_{ij} - \frac{l_{i+}}{2} - \frac{l_{+j}}{2} + \frac{l_{++}}{4} \quad (11)$$

It is known that the main effects in the loglinear μ -term model are directly related to cross-product ratios. Note that for $\mu_{1(1)}$ the terms in P appear with a positive sign whenever variable 1 is at level 1 and similarly, for $\mu_{2(1)}$ when variable 2 is at level 1. For $\mu_{12(11)}$, the positive sign appears whenever both variables are on the same level.

The model described so far applies to a table of probabilities that sum to 1. If instead we consider a table of expected counts m_{ij} that sum to a grand total of $N = \sum_{ij} m_{ij}$, we have $m_{ij} = NP_{ij}$. Hence,

$$\begin{aligned} \log m_{ij} &= \log N + \log P_{ij} \\ &= \mu' + (\mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}) \end{aligned}$$

where, $\mu' = \mu + \log N$. To compute the interaction terms the method described above is also applicable if $l_{ij} = \log P_{ij}$. It follows also that α can be defined similarly as the cross-product ratio of expected counts instead of probabilities.

This model is applicable in two sampling situations: (a) simple random sampling scheme (random sample of an entire population) and (b) a

sampling scheme where the marginals are fixed (e.g. row marginals). Two segments of the population are selected but we ignore how the distribution of individuals in the selected segments relates to the distribution of the whole population. Because of this, the μ -terms in the loglinear model have further constraints.

An example Tables 3 give the mean, main effects, interaction effects and the cell fitted values corresponding to the contingency table presented above in Table 2. Since we have a table with two dichotomous variables, the linear contrasts on the main effects (6) and interaction terms (7) hold true. The table on the right provides the fitted cell probability values which are equal to the cell $\log P_{ij}$ values. The model applied to the data is the *saturated model* in (2).

effect	score				
μ	-3.68				
$\mu_{1(1)}$	-0.6869794				
$\mu_{1(2)}$	0.6869794				
$\mu_{2(1)}$	-1.924842				
$\mu_{2(2)}$	1.924842				
$\mu_{12(11)}$	1.031224				
$\mu_{12(12)}$	-1.031224				
$\mu_{12(21)}$	-1.031224				
$\mu_{12(22)}$	1.031224				
		$cell_{ij}$	expected counts	P_{ij}	fitted
		(1,1)	2142	0.00518	-3.475892
		(1,2)	12794	0.03093	-5.951617
		(2,1)	1076	0.00260	-3.475892
		(2,2)	397575	0.96128	-0.039484

Table 3: Mean, row and column main effects and interaction effects of a 2x2 table.

3.3 Larger two-dimensional tables

Let us now define a log-linear model for a two-dimensional contingency table with $I \times J$ elementary cells, where I are the different levels of variable 1 and J the different levels of variable 2. The loglinear model defined in terms of the expected counts m_{ij} is similar to that for 2x2 contingency tables (13).

$$\sum_{i,j} m_{ij} = N \quad (12)$$

$$\log m_{ij} = \mu + \mu_{1(j)} + \mu_{2(i)} + \mu_{12(ij)} \quad (13)$$

The number of parameters contained in each μ -term is a function of I and J . The constraints on these μ -terms are unaltered (14).

$$\sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = \sum_i \mu_{12(ij)} = \sum_j \mu_{12(ij)} = 0 \quad (14)$$

for $i=1,\dots,I$ and $j=1,\dots,J$.

Computing the main effects proceeds in the same way as for the 2x2 table:

$$\begin{aligned} \mu &= \frac{l_{++}}{IJ} \\ \mu_{1(i)} &= \frac{l_{i+}}{J} - \frac{l_{++}}{IJ} \\ \mu_{2(j)} &= \frac{l_{+j}}{I} - \frac{l_{++}}{IJ} \\ \mu_{12(ij)} &= l_{ij} - \left(\frac{l_{i+}}{J} + \frac{l_{+j}}{I} \right) + \frac{l_{++}}{IJ} \end{aligned}$$

Degrees of freedom The constraints on the values of the μ -terms (14) reduce the number of independent parameters represented by each μ -term. The number of parameters for each μ -term are listed under 'degrees of freedom' because this is how parameters are viewed when fitting models to data. This 'index' is a measure of how much information the model uses to characterize the observed cell counts. The degrees of freedom are used to characterize models during model comparison. Models with fewer degrees of freedom are more interesting since they achieve a greater degree of generalization over the observed data and in sum, they are more parsimonious. The sum IJ of all parameters in the *saturated model* matches the number of elementary cells (i, j) .¹

The μ -term parameters can be redefined in terms of cross-product ratios and thus, we can obtain linear constraints (instead of logarithmic ones).

The log odds of a 2x2 table with 2 rows and j columns is:

$$\log \frac{m_{1j}}{m_{2j}} = l_{1j} - l_{2j} = 2(\mu_{1(1)} + \mu_{12(1j)})$$

Let $\alpha_{r,s}$ be the cross-product ratio for columns r and s :

$$\log \alpha_{r,s} = \left(\log \frac{m_{1r}}{m_{2r}} - \log \frac{m_{1s}}{m_{2s}} \right)$$

Taking the logarithm of the product of all $J - 1$ such α -terms yields:

$$\log(\alpha_{1,2}\alpha_{1,3}\dots\alpha_{1,J}) = 2J\mu_{12(11)}$$

¹A saturated model includes all the main effects and interaction effects. This is the model with the least predictive power because it has as many parameters as elementary cells.

In this manner, μ -term parameters are redefined in terms of cross-products and we obtain the linear constraints:

$$\mu_{12(11)} = \frac{1}{2J} \sum_{j=2}^J \log(\alpha_{1.J}) \quad (15)$$

$$\mu_{1(1)} = \sum_j \frac{l_{1j}}{J} - \sum_{ij} \frac{l_{ij}}{2J} = \sum_j \left(\frac{l_{1j} - l_{2j}}{2J} \right) \quad (16)$$

$$\mu_{2(1)} = \frac{J-1}{2J} (l_{11} + l_{21}) - \sum_{j=2}^J \frac{(l_{1j} + l_{2j})}{2J} \quad (17)$$

This means that to express such parameters we can (i) use the definitions of μ -terms as linear contrasts of the l_{ij} shown above (15–17) or (ii) give the corresponding multiplicative form (as $e^{\mu_{1(1)}}$, $e^{\mu_{2(1)}}$ or $e^{\mu_{12(1j)}}$).²

3.4 Model for three-dimensional tables

First, a few remarks on notation. The notation used above to describe cells of a two-dimensional array (table) is extended to multiway tables by adding more subscripts. The number of subscripts normally matches the number of variables, but exceptions occur. Subscripts match the dimension of a particular arrangement of the cells. The probability of a count falling in cell (i, j, k) is P_{ijk} and the expected count is m_{ijk} , where $i = 1 \dots I$, $j = 1 \dots J$ and $k = 1 \dots K$. A complete three-dimensional array of size $I \times J \times K$ has IJK cells.

The single loglinear model for the whole $I \times J \times K$ array is described in equation (18). The subscripted μ -terms preserve the property that their sum equals zero. For the parameters of each μ -term there is one absolute value in the $2 \times 2 \times 2$ table. In this table, each μ -term contributes one degree of freedom. In the $I \times J \times K$ table, it is possible that some of the μ -term parameters equals zero while other parameter has a large value. This is a case where each μ -term may contribute more than one degree of freedom.

$$l_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)} \quad (18)$$

Interpretation of the parameters All **two-factor effects** $\mu_{12(ij)}$, $\mu_{13(ik)}$ and $\mu_{23(jk)}$ can be interpreted as deviations from the overall effect of a single variable. The **three-factor effect** $\mu_{123(ijk)}$ corresponds to the difference between the average value of $\mu_{12(ij)}$ across tables ($2 \times 2 \times k$) and the particular value exhibited by table k .

²The multiplicative form of the μ -term parameters can be found in Bishop et al. (1984) among other references.

Linear contrasts Similarly as for the 2x2 and 2xJ table, every μ -term can be rewritten as a linear contrast of the logarithm of the four cells or as cross-product ratios.

Degrees of freedom Table 4 enumerates the number of independent parameters in the model contributed by each of the main effects and interaction effects.

Number in level	level	degrees of freedom
1	overall mean	1
3	one-factor terms	(I-1)+(J-1)+(K-1)
3	two-factor terms	(I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1)
1	three-factor term	(I-1)(J-1)(K-1)
Total		IJK

Table 4: Number of degrees of freedom at each level in a saturated loglinear model for a IxJxK array.

Interpretation of models The model and linear contrasts above are applicable to a random sampling scheme. The *saturated* loglinear model is the most complex model and it fits the data perfectly. To find a simpler *unsaturated* model that shows a good fit (to the data), one removes interaction effects *gradually* until a decent fit is found. This means proceeding further with the deletion of the three-factor interaction term, and then (some or all) two-factor interaction terms. If the variables are independent of each other, removing the two-factor and three-factor interaction terms should not affect the model's goodness of fit. If in the model the interaction term effects between (any of) the variables are assigned a zero value, the variables are completely independent; this implies that the *model of independence* holds true:

$$l_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)}$$

3.5 Parameter estimation

The parameters of the μ -terms in log-linear models can be estimated in several ways: using direct estimates or maximum likelihood estimates (MLEs). Although MLEs are the most commonly used Bishop et al. (1984) direct estimates are occasionally possible.

A rectangular *complete* table has a nonzero probability of a count occurring in every cell. Under the *model of independence*, the cell expected values coincide with the maximum likelihood estimates (19).

$$\hat{m}_{ij} = \frac{x_{i+}x_{+j}}{x_{++}} \quad (19)$$

where $x_{i+} = \sum_j x_{ij}$, $x_{+j} = \sum_i x_{ij}$ and $x_{++} = \sum_{ij} x_{ij}$

That is, this model has direct estimates. In general, we encounter models that have no direct estimates. In this case, maximum likelihood estimation is performed with iterative methods like *Iterative Proportional Fitting* or the *Newton-Raphson algorithm*. We only mention these iterative methods here. Some statistical packages allow the user to choose estimation method. For further details on estimation methods refer to Bishop et al. (1984) or Agresti (2002).

3.6 Evaluation procedure

Whatever our purpose, we want to find the model that best fits our data. A search procedure is followed that compares how well different models fit our data. The goodness of fit of a model is typically measured with the Pearson's X^2 statistic or with the log-likelihood ratio G^2 . (Other measures exist.) Small values of these statistics prove model fit. To establish why a model fails to fit our data, exploration of the Pearson's residuals helps to identify effects that are not taken into account by a model. This search procedure is often available in most statistical packages.

3.7 Loglinear models and MWE identification

Applied to language models, categorical variables can have words as values. The models can be used to identify interactions between words that can be part of a multiword expression. An example of this has been shown in section 3.2. Indirectly, the models are applied to identify expressions made up of words that exhibit a strong interaction between them.

Loglinear models has been used to identify multiword expressions earlier. Blaheta and Johnson (2001) approximated the lexical association between the component words of English phrasal verbs (verb and particles) as the association parameter values of λ terms in a saturated loglinear model. The λ terms are equivalent to the μ terms in our description above. We applied Blaheta and Johnson's implementation of log-linear models to identify support verb constructions in a Dutch corpus. Although the models reached a rather high precision, the recall was not satisfactory probably due to data sparseness (Villada Moirón 2005, chp.5). Nowadays, we have access to a larger quantity of annotated data, thus, we are doing further experiments with similar techniques.

4 Latent semantic analysis

Latent semantic analysis (LSA) is a machine learning model that induces representations of the meaning of words by analyzing the relation between words and passages in large bodies of text. LSA was originally developed in the context of information retrieval (Deerwester, Dumais, Furnas, Landauer, and Harshman, 1991) as a way of overcoming problems with polysemy and synonymy. In natural language processing, LSA has been used in a variety of tasks such as recognising synonymy, word sense disambiguation, document-term similarity and even finding correct translations in parallel corpora.

4.1 Classifying multiword expressions using LSA

(Baldwin, Bannard, Tanaka and Widdows 2003) used LSA to determine the (meaning) similarity between a multiword expression (MWE) and its constituent words. Their objective was to classify MWEs from a semantic point of view, namely, demarcate single decomposable MWEs (*traffic light*) from idiosyncratic decomposable (*spill the beans*) and non-decomposable MWEs (*shoot the breeze*).

LSA is used to build a vector space model in which term-term similarities could be measured. A word is represented as a vector that enumerates all the co-occurring content words observed within the same sentence in a corpus. They proceed as follows: (i) Select 1000 most frequent words in a corpus. These terms are used as column labels in a matrix. (Aiming at better context vectors, they use a stoplist with prepositions, determiners, etc.) (ii) Select 50,000 most frequent terms in corpus and assign them to row labels. The rows can be thought of as word-vectors. Count the number of times each term co-occurs with a content-bearing word within the same sentence. As terms they include noun noun compounds and particle verbs. (iii) To avoid data sparseness, a singular value decomposition (SVD) algorithm³ reduces the number of dimensions from 1000 to 100. (iv) Measure the similarity between two vectors by using the cosine of the angle between them.

The corpus has POS-tags. Terms include their part of speech so that the vector describing e.g. *fire_{noun}* will be a different vector from *fire_{verb}*. This syntactic distinction avoids clustering two different uses of a word into one single vector Baldwin et al. (2003). Two different models were built, one for noun-noun compounds and one for particle verbs. The models compare a MWE to the constituent words. The neighbors of the MWE are compared to the neighbors of the constituent words. As an example, the MWEs *cut off* and *cut out* are compared to the single verb *cut*. The cosine similarities $\text{sim}(\text{cut}, \text{cut_out})$ and $\text{sim}(\text{cut}, \text{cut_off})$ are 0.433 and 0.183, respectively. This similarity score reflects the fact that the terms (cut, cut out)

³Each word is projected onto the n-dimensional subspace which gives the best least-squares approximation to the original data.

share similar meaning. Then, on the basis of the similarity score, one may conclude that *cut out* is a semantic decomposable MWE because its context is close to the context of the singleton verb *cut*.

The idea underlying this approach is a very intuitive one. It exploits a property of natural language that words with similar meaning tend to occur together.⁴ LSA appears to be well-suited for semantic clustering and word sense disambiguation Widdows, Dorow and Chan (2002). However, in the task at stake, the LSA model does not perform very well in practice. Surprising result: the LSA model performs better over low-frequency items. A possible cause for this is the fact that high-frequency items are more polysemous Baldwin et al. (2003). Schone and Jurafsky (2001) also used LSA to improve the identification of multiword units by trying to eliminate proposed MWUS which are semantically compositional. Their performance results had also been reported to be rather poor Schone and Jurafsky (2001).

5 Software to apply these models

Exploring dependencies between two or three variables represented in a small contingency table can be done with the R statistical package. R is a programming language for statistical analysis. For large contingency tables with multiple dimensions and multiple levels per dimension, R loglinear models are difficult to apply. The algorithm used for the loglinear models during estimation of the association parameters seems to be computationally very demanding and R fails. Instead, we decided to use TADM. The Toolkit for Advanced Discriminative Modeling (TADM) is a C++ implementation for estimating the parameters of discriminative models, such as maximum entropy models. This toolkit is freely available at <http://tadm.sourceforge.net/>.

Once we have a reasonable list of MWEs, we will apply latent semantic analysis to classify the expressions from a semantic perspective. Many free resources are available to apply this technique. We will experiment by using SVDLIBC <http://tedlab.mit.edu/~dr/SVDLIBC/>, a tool to perform singular value decomposition on large matrices of data.

6 Case study: Quantifying linguistic properties

In section 1.1 we enumerated a few general characteristics of multiword expressions, how can we capture the mentioned theoretical linguistic properties of MWEs and use them to solve the identification problem? To what extent do these 'properties' identify MWEs? Can a simple loglinear model identify a dependence between them?

⁴In work on semantic classification of verbs, it is also known that LSA has trouble in classifying words that share an antonym relationship (sell/buy) due to similarity in their contexts.

6.1 MWE candidates extraction

Rather large automatically annotated corpora of Dutch are available. With the Alpino parser⁵, parts of the *Twente Nieuws Corpus* (TWNC) and CLEF data used to evaluate question answering systems was annotated. With a query over these corpora, we extracted all possible realizations of the syntactic pattern VERB + PP. We limit the search to documents that contain one of a set of 12 verbs (most of them have a support verb usage). These verbs are: *brenge*, *doen*, *gaan*, *geven*, *hebben*, *houden*, *komen*, *krijgen*, *maken*, *nemen*, *stellen* and *zitten*. Other verbs present in the selected documents are also tallied and end up in our dataset. While coding the candidates, NPs are reduced to the head noun’s lemma and verbs are lemmatised, too. Other arguments inside the verb phrase node are ignored. A sample of more than 160,000 triples was collected. Table 5 shows an excerpt from the dataset.

candidate	frequency
uit familie kom	70
uit gezin kom	74
uit hand geef	153
uit hoek kom	183
uit land kom	239
uit lucht val	182
uit mond kom	94
uit onderzoek kom	142
uit onderzoek blijk	372
vanaf er kom	86
van belang ben	595

Table 5: Candidate PREP NOUN VERB triples and their frequency extracted from annotated corpora.

6.2 Quantifying linguistic properties

Among the theoretical linguistic properties, we consider a lexical property and a trend in the syntactic behavior. These are operationalized by measuring: (i) lexical affinity between words, (ii) head dependence between component words in the candidate triple and (iii) the likelihood that the (potential) argument component PP occupies preverbal position in the Dutch verbal cluster.

Lexical affinity Two different measurements are used. On the one hand, association measures such as salience and Dunning’s log-likelihood. With

⁵Available at <http://odur.let.rug.nl/~vannoord/alp>

V1sta:V2onder:V3druk	4.34120e+00
V1lig:V2voor:V3hand	4.03183e+00
V1kom:V2tot:V3stand	3.98308e+00
V1ga:V2van:V3start	3.96905e+00
V1sta:V2op:V3programma	3.93207e+00
V1breng:V2om:V3leven	3.73351e+00
V1kom:V2om:V3leven	3.72854e+00
V1lig:V2onder:V3vuur	3.51669e+00
V1stel:V2aan:V3orde	3.40583e+00

Figure 1: Log-linear model applied to a $IxJxK$ table that cross-tabulates verbs, prepositions and nouns. (Association) Parameter values of the $\mu_{123(ijk)}$ interaction terms.

V1kom:V2over:V3brug	2.47688e+00
V1zit:V2op:V3huid	2.46116e+00
V1heb:V2onder:V3knie	2.45472e+00
V1help:V2aan:V3baan	2.02373e+00
V1ga:V2uit:V3eten	1.79986e+00
V1breng:V2naar:V3voren	1.14369e+00
V1vertrek:V2op:V3tijd	6.97125e-01

Figure 2: Log-linear model applied to a $IxJxK$ table that cross-tabulates verbs, prepositions and nouns. Some candidates assigned lower association parameter values of the $\mu_{123(ijk)}$ terms.

these tests we measure the statistical association between the component words in the (P, N, V) candidates treated as bigrams (P_N, V) . On the other hand, we compute the association parameters of a multi-way $IxJxK$ array that cross-classifies three multivalued variables: **preposition**, **noun** and **verb**. Figures 1 and 2 show the parameter values of the $\mu_{123(ijk)}$ interaction terms.

Head dependence measures the number of verbs that select for a given PP. Merlo and Leybold (2001) used the head dependence as a diagnostic to determine the argument (or adjunct) status of a PP. PP-arguments should only appear with a group of heads that lexically select them, while PP-modifiers can appear with a greater range of different heads. Baldwin (2005) also suggests that this measure ought to be helpful to identify prepositional verbs. In our data, a PP selected by a few verbs is likely to be either a regular prepositional complement or a fixed argument in a MWE. We use two variants of head dependence: (a) score measured in integers and (b) score measured as the amount of entropy observed among the co-occurring verbs for a given PP (20). The integer score and the entropy essentially measure the same 'head dependence' but in a different way. The entropy of a tuple $H(P, N)$ is measured over all its instances (tokens) taking into

van wijs breng	142	1	0
aan pols houd	87	2	0,0529733
bij paaltje kom	83	2	0,0645814
in hongerstaking ga	89	2	0,061047
in maling neem	71	2	0,0731901
met keukenpapier maak	148	2	0,0402724
met vut ga	70	2	0,0740226
om geval ga	84	2	0,0639617
om groep ga	102	2	0,0546588
om miljard ga	80	2	0,0665216
tot ontploffing breng	164	2	0,477524
uit gezin kom	74	2	0,121692
in cassatie ga	102	3	0,10836
op vuist ga	80	3	0,178498
te dood breng	74	3	0,139933
tegenover daar sta	159	3	0,752765
tegenover daar stel	106	3	0,752765
tot akkoord kom	215	3	0,16812
tot uitbarsting kom	99	3	0,110993
tot vergelijk kom	115	3	0,219639
aan beurt kom	157	4	0,229696
aan kook breng	302	4	0,520165
op adem kom	89	4	0,222116

Figure 3: Measuring head dependence. Each candidate pattern (P, N, V) shows the frequency of the pattern, the number of verbs co-occurring with the (P, N) and the entropy of the (P, N) .

account all co-occurring verbs.

$$H(P, N)_i = - \sum_i \sum_j \frac{f((P, N)_i, V_j)}{f(P, N)_i} \log \frac{f((P, N)_i, V_j)}{f(P, N)_i} \quad (20)$$

Figure 3 displays the candidate patterns whose component PP co-occurs with only a small number of verbs (in our sample). The frequency, number of co-occurring verbs and entropy are given. As an example, *aan pols houden* was found 87 times; the PP *aan pols* is found with two co-occurring verbs. Since only two verbs select for this PP the entropy score is very low.

The head dependence measured as entropy is more informative and therefore more accurate in approximating the lexical affinity between the words. If we were to rely on the head dependence of a PP measured as the number of co-occurring verbs, *tegenover daar staan* is selected by 3 verbs, whereas *op adem komen* is selected by 4 verbs. The integer score

suggests that **tegenover daar staan** is a better candidate for MWE. The entropy score confirms the opposite: **op adem komen** is a better candidate.

Position of PP argument in verb final contexts Knowledge of how language works tells us that there are obligatory syntactic dependents and optional dependents. Among obligatory: productive complements, fixed arguments in MWEs and predicative phrases and among optional: adjunct modifiers. Fuzzy cases exist which are hard to classify. Broekhuis (2004) studied the necessary and sufficient conditions to label a PP as a complement or as an adjunct. Among the syntactic tests to distinguish complements from adverbial PPs, Broekhuis (2004, pg.107) argued that the PP complements are closer to the verb in verb final context than adverbial PPs. Jack Hoeksema (p.c.) further maintains that fixed arguments in fixed expressions show a preference for the position immediately preceding the verbal cluster.

Starting from these observations, we register the position in which a PP is observed in a sentence. The position values we chose are: immediately preceding the verb (**ipr**), before the verb (**pr**) and following the verb (**fol**). In this first case study, we only use the distinction between **ipr** and some other position (**opos**) for practical reasons. For a given PP VERB type, we compute the probability of seeing the PP in the **ipr** position (that is immediately preceding the verb cluster) in a verb final context. Figure 4 shows the relevant measurements.

6.3 Are the quantified properties discriminating?

Our expectations are:

1. those (P, N, V) patterns with a high lexical affinity score are more likely to be MWES,
2. those (P, N, V) patterns with a very low score of head dependence and a very low entropy are more likely to be MWES
3. those (P, N, V) patterns in which the PP shows a strong preference to immediately precede the verb in verb-final contexts are more likely to be MWES.

We ranked the candidate triples on the basis of each measured property. The resulting rankings list many MWES among the higher ranks and many regular phrases among the lowest ranks thus, to some extent confirming our expectations. Table 6 summarizes these findings.

To some extent, the chosen empirical linguistic properties are able to discriminate one type of expressions from the other. If we were to use this model to identify MWES, a threshold applied on the magnitude of each

aan_afspraak|houd|vfi|132|251|0.88
aan_bak|kom|vfi|74|123|0.99
aan_bod|kom|vfi|310|813|1.00
aan_hoofd|sta|vfi|50|70|0.68
aan_kaak|stel|vfi|351|415|1.00
aan_kook|breng|vfi|16|302|1.00
aan_licht|breng|vfi|209|344|1.00
aan_licht|kom|vfi|358|782|1.00
aan_orde|stel|vfi|501|601|1.00
aan_pols|houd|vfi|53|87|0.92
aan_regel|houd|vfi|172|255|0.89
aan_slag|ga|vfi|377|651|1.00
aan_slag|kom|vfi|125|191|1.00
aan_woord|kom|vfi|144|340|1.00
in_bezit|kom|vfi|65|118|0.80
in_bloed|zit|vfi|29|113|0.97
in_boek|kom|vfi|48|93|0.40
in_botsing|kom|vfi|86|194|0.99
in_brand|steek|vfi|60|79|1.00
in_brief|schrijf|vfi|75|134|0.25
in_gevangenis|zit|vfi|195|357|0.95
in_geweer|kom|vfi|59|115|1.00
in_greep|houd|vfi|95|140|0.99
in_groep|zit|vfi|47|94|0.66
in_hand|ben|vfi|69|99|0.90

Figure 4: Each candidate expression enumerates the frequency of the PP occurring in a verb final context, the frequency of the candidate (P, N, V) and the probability of seeing the PP in *ipr* position.

Attribute	magnitude	MWE expression
lexical affinity	high	✓
head dependence (integer)	small	✓
head dep. entropy	small	✓
'ipr' position in VC	prob. close to 1	✓

Table 6: Selecting empirical linguistic properties to delimit the landscape of MWES.

linguistic property (attribute) needs to be found out empirically. This approach has proved useful in identifying determinerless PPs in Dutch by van der Beek (2005). Such a threshold would establish the values **high**, **small** and eventually, set apart true positives (MWES) from false positives (regular phrases).

However, no ranking is sufficiently good. That means, no single empirical property on its own is sufficiently discriminative as to delimit the subspace of multiword expressions M . This suggests a more expressive model that incorporates more than one property of MWES. Two models emerge: (i) a combined ranking system of all the measurements and (ii) a generalized linear model that incorporates the empirical linguistic properties as independent variables. Latest research shows that a ranking system that combines several empirical properties produces better precision and recall (Baldwin 2005). We will evaluate such a model against a generalized linear model.

6.4 Measuring dependencies with a loglinear model

Generalized linear models (thus also loglinear models) can take into account complex interactions between variables, provided there is enough data. During model fitting one might want to know whether the relationship between two variables is of any interest. Applying a loglinear model to the contingency table that cross-tabulates (only) the two variables might reveal significant interaction effects between the variables at stake.

For illustration, let us ask ourselves whether expressions that show a tendency to precede the verb inside the verb cluster, also show a low head dependence entropy score (thus, a strong attraction between the PP and the verb). We took the same sample of VERB PP instances described earlier in section 6.1 and cross-tabulated the head dependence entropy with the position in the verb cluster as displayed in Table 7.⁶

head dependence	Position in verb cluster	
	f_ipr	f_opos
low	58163	42701
medium	13625	21607
high	7853	16616

Table 7: Cross-tabulation of the position of a PP phrase inside the verb cluster and the value of the head dependence between the PP and its selecting verbs.

⁶Using the head dependence entropy score, we assigned each (P, N, V) triple a value within a scale between **low** and **high** depending on the entropy score to reduce the category value levels.

With the R statistical package,⁷ we tried to find a loglinear model that fits the data in the table. After two iterations, the best fitting model for our data was found. Refer to Figure 5 to see the results.

Among the information returned by this software, we get two goodness of fit measures: the log-likelihood ratio G^2 and Pearson's χ^2 . For a contingency table with two degrees of freedom, both measures exhibit rather big values which suggests that the model fit is far from perfect. This lack of fit can also be observed in the estimated cell values shown under `$fit`. They diverge quite a lot from the observed cell values in Table 7 above. Parameters of the row and column effects are also given. By default, the R `loglin()` function uses the *iterative proportional estimation* algorithm to estimate the parameter values. Summing up the parameters of each main effect, we see that the linear constraint holds: they add up to zero. Two figures show that a significant dependence may be at stake between the two factors: the parameter under `low` of the row effect μ_1 and the intercept parameter of the interaction term (μ_{12}). If the two factors (position in verb cluster and head dependence) were completely independent this intercept parameter should have been equal to zero. The parameter values of the row effect show that: there is a small preference for `low` over other categories of the factor `head dependence`. This observation has to be interpreted carefully since this effect might be a side-effect of the method we used to recode the ordinal variable head dependence entropy into nominal values (`low,medium,high`). However, the interesting fact is that among triples with `low` head dependence a larger number of PP arguments immediately precede the verb in a verb final context. This pattern is exactly the opposite of the pattern shown by the head dependence `medium` and `high` levels. The conclusion we arrive at: there seems to be a statistical dependence between the variables, thus, we can proceed to incorporate an interaction effect between the two variables in a complex loglinear model.

6.5 Summary

This explorative case study described some necessary tasks previous to applying loglinear models. These tasks were: (i) MWE candidates extraction, (ii) quantifying empirical linguistic properties and (iii) establishing discriminating properties. We also gave an example of using the R statistical package to measure dependencies between properties with a loglinear model.

Some observations:

- Extraction of candidate patterns and extraction of evidence to approximate theoretical linguistic properties of MWEs can be done from (automatically) annotated corpora with syntactic information.

⁷Available at <http://www.r-project.org/>.

```

> loglin(vps.hd_pos0,c(1,2), fit=TRUE, param=TRUE)
2 iterations: deviation 1.455192e-11
$lrt
[1] 7402.48

$pearson
[1] 7307.012

$df
[1] 2

$margin
[1] 1 2

$fit
      f_ipr  f_opos
low    50029.97 50835.01
medium 17475.44 17756.64
high   12137.06 12332.35

$param
$param$(Intercept)"
[1] 10.00563

$param$"1"
      low      medium      high
0.8227283 -0.2290975 -0.5936308

$param$"2"
      f_ipr      f_opos
-0.007981475  0.007981475

```

Figure 5: Results of fitting a loglinear model to a 2x3 contingency table using R.

- Theoretical linguistic properties such as lexical affinity can be empirically approximated with association measures (log-likelihood or salience), a loglinear model estimated parameters and the head dependence measured as entropy. The head dependence measured as entropy is more informative and therefore more accurate. The preferred syntactic location in a verb final context can be captured as the probability of a PREP NOUN of being seen in `ipr` position.
- None of the selected empirical linguistic properties on its own is a good discriminating attribute of MWES. A combined ranking system surely increases precision and recall (cf. (Baldwin 2005)). We will evaluate such a model against a generalized linear model.

References

- Agresti, A.(2002), *Categorical Data Analysis*, John Wiley and Sons, New York.
- Baldwin, T.(2005), Looking for prepositional verbs in corpus data, *Proc. of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in computational linguistics formalisms and applications*, Colchester, UK.
- Baldwin, T., Bannard, C., Tanaka, T. and Widdows, D.(2003), An Empirical Model of Multiword Expressions Decomposability, *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 89–96.
- Bishop, Y., Fienberg, S. and Holland, P.(1984), *Discrete Multivariate Analysis. Theory and Practice*, 8th edn, MIT Press, Cambridge,Massachusetts.
- Blaheta, D. and Johnson, M.(2001), Unsupervised learning of multi-word verbs, *39th Annual Meeting and 10th Conference of the European chapter of the Association for Computational Linguistics (ACL39)*, CNRS, Toulouse, France, pp. 54–60.
- Broekhuis, H.(2004), Het voorzetselvoorwerp, *Nederlandse Taalkunde* **9**(2), 97—131.
- Merlo, P. and Leybold, M.(2001), Automatic distinction of arguments and modifiers: the case of prepositional phrases, *Procs of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, Toulouse. France, pp. 121–128.
- Paolillo, J.(2002), *Analyzing Linguistic variation. Statistical models and methods*, CSLI Publications.

- Schone, P. and Jurafsky, D.(2001), Is knowledge-free induction of multiword unit dictionary headwords a solved problem?, *Proc. of the 2001 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, pp. 100–108.
- van der Beek, L.(2005), *Topics in Corpus-based Dutch Syntax*, PhD thesis, Alfa-Informatica, University of Groningen, The Netherlands.
- Villada Moirón, B.(2005), *Data-driven Identification of fixed expressions and their modifiability*, PhD thesis, University of Groningen.
- Widdows, D., Dorow, B. and Chan, C.-K.(2002), Using parallel corpora to enrich multilingual lexical resources, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 3)*, Las Palmas, pp. 240–247.