Syracuse University

SURFACE

School of Information Studies - Faculty Scholarship

School of Information Studies (iSchool)

July 1992

Discourse-level Analysis of Abstracts for Information Retrieval: A Probabilistic Approach

Robert N. Oddy
Syracuse University

Recommended Citation

Oddy, Robert N., "Discourse-level Analysis of Abstracts for Information Retrieval: A Probabilistic Approach" (1992). *School of Information Studies - Faculty Scholarship*. 155.
https://surface.syr.edu/istpub/155

Discourse-level Analysis of Abstracts for Information Retrieval: A Probabilistic Approach

Robert N. Oddy

The objective of this research is to contribute to our knowledge of how people seek information, and how computer systems can be designed to help in this process. Most information retrieval research since the field emerged in the 1950's has reduced these questions to that of trying to determine how documents relevant to a user's query might be selected from a large collection of texts---a question that has proved remarkably difficult to answer. The present work takes the stance that this particular reduction increasingly limits progress towards the objective stated above. It is directed instead towards the development of a framework for IR based on the notions of discourse and human communication.

1. Information Retrieval and Discourse

An argument suggesting the use of discourse-level linguistic analysis of texts for information retrieval purposes was given in a previous paper [1].

> "Much of human discourse is concerned with sharing experience of the situations with which we must cope. It is directed towards improving our ability to recognize common situations, and to respond effectively to them. Specifically, document abstracts are written with the intention of informing people who belong to the same community as the authors and are engaged in similar work. They should, therefore, be able to recognise common situations. The situations of interest to an author are discernable in the discourse-level structure of an abstract." [1, p124]

Evidence for these claims was presented in the paper, as was some preliminary work on automatic discourse-level text analysis and implications for system design. It was argued that the next step should be the development of a prototype IR system which makes use of discourse-level structures to allow users to express situation-related aspects of their information needs. This is prerequisite to performing empirical work on our hypotheses to do with the role of situational information and discourse structure in IR. [1, pp166-7]

2. Discourse-level structure of empirical abstracts

This work makes use of a hierarchical, componential text structure for empirical abstracts proposed and thoroughly investigated by Liddy[2]. The hierarchy is displayed in Figure 1, and an example of an analysed abstract is given in Figure 2. There are 37 component types in the whole hierarchy (known as the Elaborated model), although one would not expect to see all of them in any particular abstract. Also, the hierarchy reflects the logical relationships which are most obvious to abstractors and readers, and although abstracts frequently follow this scheme, components are displaced in a significant number of specific texts. Liddy suggested two sub-structures, on the basis of increasing typicality in abstracts and abstractors' perceptions: the Typical model (15 component types) and the Prototypical model (7 component types). These are also shown in Figure 1.

In the course of her Ph.D dissertation work, Liddy developed a small corpus of empirical abstracts, analysed into components (Elaborated model), by hand [2]. These abstracts were subsequently coded with a simple bracketing scheme to indicate the positions of the components in the texts (Figure 3). In this corpus, abstracts from two online databases are represented:

> from ERIC (education): 150 abstracts containing 1754 components
> from PsycINFO (psychology): 126 abstracts containing 1464 components

3. Discourse analysis: a Probabilistic approach

A perusal of typical abstracts leads to the impression that an analysis based on linguistic processes (syntactic and semantic analysis) would be complex, and involve high computation costs. On the other hand, it has been clear since Liddy did the initial analysis that there are clues in the texts which might provide probabilistic evidence of the discourse-level structure. This should lead us to relatively fast computer programs which could be applied to large databases. A very primitive probabilistic model was described in [1], and this work is considerably extended in the present paper.

In this work, the aim of a probabilistic analysis of texts into their discourse-level structure is taken to be:

> *the assignment of text fragments to the discourse-level components of the text-type, with associated estimates of the probability of correct assignment.*

From these assignments, using a decision rule, one or more structural analyses of the whole text can be proposed with varying degrees of certainty. However, this step is beyond the scope of the present paper.

The probabilities can be estimated by combining evidence from clues observed in the text once statistical information has been derived from a corpus of typical texts. The information used here for the estimation is as follows:

- (i) Relative frequency of occurrence of the various component types in a corpus of typical texts;
- (ii) Frequency distributions of lexical clues (words, stems, word classes) in the components within the corpus;
- (iii) Structural information; specifically the frequency distribution of the adjacency of all possible pairs of component types in the corpus.

There is another significant aspect to the problem, namely how are the text fragments to be obtained in the first place? It was decided to treat this problem separately at first, even though it is unlikely that in a successful system fragmentation can be performed independently of the estimation of probabilities. So, the research strategy was to explore first the probability estimation, assuming that the fragmentation could be done correctly (using test data that had been fragmented by hand). Then some quite simple automatic fragmentation methods were explored.

4. The Probabilistic Models

4.1 Lexical clues

The model for the use of lexical clues is a simple application of Bayes' theorem.

Suppose a text fragment, F, contains a set of clues, S. We can write this as follows:

> F is represented by the clue vector $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$, where n is the size of the clue vocabulary,
>
> and $x_i = 1$, if clue i is present in S, otherwise $x_i = 0$.

For each component type, c ($1 \le c \le C$), we want $P(c \mid \boldsymbol{X})$, *i.e.* the probability that the fragment F is a component of type c, given that clue vector $\boldsymbol{X}$ is observed. Invoking Bayes' theorem (and assuming clues occur independently):

$$P(c \mid \boldsymbol{X}) = \frac{P(c) \times P(\boldsymbol{X} \mid c)}{P(\boldsymbol{X})}$$

$$= \frac{P(c) \times \prod_{i=1}^{n} P(x_i \mid c)}{\sum_{j=1}^{C} \{P(j) \times P(\boldsymbol{X} \mid j)\}} \qquad (1)$$

From a typical, previously analyzed corpus of abstracts, the parts of this expression can be estimated as follows:

$$P(c) = N_c / T$$

$$P(x_i = 1 \mid c) = \frac{\max(n_{ci}, 0.5)}{\max(N_c, 0.5)}$$

$$P(x_i = 0 \mid c) = 1 - P(x_i = 1 \mid c)$$

where N_c = no. of components of type c
n_{ci} = no. of components of type c containing clue i
T = total no. of components

The denominator of (1) is the sum of the numerators (over c).

A clue is a word (stem, or semantic class, for instance) whose occurrence in a text fragment can tell us something about the component type, *i.e.* one whose frequency distribution in components is not random.

Finding a set of clues, using a small corpus is difficult, and was the subject of quite extensive work early in the present project. This process will be described in the next section.

4.2 Choice of lexical clues

Potential clues are extracted from the corpus, and may include:

(i) single whole words (no exception list)
(ii) stems (optional), where different from the original words
(iii) class names (optional), for those words which are found to belong to predefined classes

For a potential clue, *i*, and component type, *c*:

n_{ci} = number of components of type *c* containing *i*;
n = number of different types of component containing *i*;
$f = \sum_{c=1}^{C} n_{ci}$;
N_c = number of instances of component type *c*.

Firstly, a simple filter is applied. Potential clue *i* is rejected if

(i) $n_{ci} \le A$ for all *c*, or
(ii) $f \le B$.

The values $A = 1$ and $B = 2$ were found to be satisfactory (sample corpus size may be a factor here). The rationale for applying the filter is that with so little data statistical information about these words is extremely unreliable.

Now, a score, S_{ci}, is computed for the association of the potential clue, *i*, with each component type, *c*, in which it occurs. (Note that a potential clue may have more than one score.)

$$S_{ci} = (1 - \frac{1}{f}).(1 - \frac{1}{N_c}).\frac{n_{ci}}{N_c}.\frac{n_{ci}}{f}.\frac{1}{n^2}$$

The first factor asymptotically approaches 1 from below as *f* increases, and serves as a way of reducing the score for the less frequent words, on the grounds that statistical errors will be greater for them. The second factor has a similar purpose with respect to less frequent component types. The other factors reflect the following criteria for a good clue:

(i) a large proportion of the components of type *c* should contain it (for some *c*)
(ii) a large proportion of its occurrences should be concentrated in one type of component
(iii) it should occur in relatively few different component types

One might expect words whose distribution across component type was highly skewed to have these properties. However, when the skew was tried as a means of selecting clues, performance (in terms of the success of the probability estimations) was inferior to that obtained with S_{ci}.

Finally, the association scores are ranked and words selected from the top until a specified number K of different clues have been found.

4.3 Use of structural evidence

Consider an abstract consisting of N fragments, f_n ($1 \leq n \leq N$). As a first attempt to capture the structural relationships between components, in a probabilistic way, we expand the probability $P(f_n = c_i)$, that fragment f_n is a component of type c_i ($1 \leq c_i \leq C$), in terms of the probability distribution for the preceding fragment:

$$P(f_n = c_i) = \sum_{j=0}^{C} \left\{ P(f_n = c_i \mid f_{n-1} = c_j).P(f_{n-1} = c_j) \right\} \qquad (2)$$

Now, we already have an estimate of $P(f_n = c_i)$ based on clue data, which we do not wish to discard, so we need to *revise* the probabilities $P(f_n = c_i)$. This can be done in a sequential manner, working forwards through the abstract, using Jeffrey's rule of conditioning:

$$P^*(f_n = c_i) = \sum_{j=0}^{C} \left\{ P(f_n = c_i \mid f_{n-1} = c_j).P^*(f_{n-1} = c_j) \right\} \qquad (3)$$

$P^*()$ denotes revised probabilities. This step depends on the assumption that the conditional probabilities do not change while the revision is taking place, which seems reasonable in this application because they express quite stable structural properties of the text-type.

In these formulae, f_0 is the (virtual) fragment preceding the first in the abstract, interpreted as the beginning of the abstract, and c_0 is its notional (fixed) component type.

Thus $P(f_n = c_0) = P^*(f_n = c_0)$ = 1, if $n = 0$
= 0, otherwise

and $P(f_0 = c_i) = P^*(f_0 = c_i)$ = 1, if $i = 0$
= 0, otherwise

Now, inverting the conditional probabilities, we can express the revision for fragment f_n in terms of the revisions that have just taken place for the preceding fragment, and knowledge about the probabilities of component sequences in abstracts.

$$\begin{aligned} P^*(f_n = c_i) &= \sum_{j=0}^{C} \left\{ \frac{P(f_{n-1} = c_j \mid f_n = c_i).P(f_n = c_i)}{P(f_{n-1} = c_j)}.P^*(f_{n-1} = c_j) \right\} \\ &= \left[\sum_{j=0}^{C} \left\{ q_{ij}.\frac{P^*(f_{n-1} = c_j)}{P(f_{n-1} = c_j)} \right\} \right].P(f_n = c_i) \qquad (4) \end{aligned}$$

where $q_{ij} = P(f_{n-1} = c_j \mid f_n = c_i)$

The initial value of $P(f_n = c_i)$ could be the one derived from lexical clues, or an estimate based on relative frequency of component types, or a constant over all *i*, indicating no prior belief.

To estimate $P^*(f_n = c_i)$, we use an estimate of q_{ij}:

$$q_{ij}est = \frac{seq(j,i)}{\min(N_i, 0.5)}$$

where $seq(j,i)$ = number of occurrences of components of type *j* followed by components of type *i*, in a sample of abstracts. In particular, $seq(0,i)$ = number of occurrences of components of type *i* occurring at the beginning of abstracts.

So, to calculate *P*()*, we have two cases:

Case (i): $n = 1$

$$P^*(f_1 = c_i) = q_{i0}.P(f_1 = c_i) = \frac{seq(0,i)}{\min(N_i, 0.5)}.P(f_1 = c_i)$$

Case (ii): $n > 1$

$$P^*(f_n = c_i) = \left[\sum_{j=0}^{C} \left\{ seq(j,i).\frac{P^*(f_{n-1} = c_j)}{P(f_{n-1} = c_j)} \right\} \right].\frac{P(f_n = c_i)}{\min(N_i, 0.5)}$$

Another possible revision procedure is to work through the abstract from the end towards to beginning. The mathematics is analogous to forward revision. So, corresponding to equation (4) is equation (5):

$$P^*(f_n = c_i) = \left[\sum_{j=0}^{C} \left\{ p_{ij}.\frac{P^*(f_{n+1} = c_j)}{P(f_{n+1} = c_j)} \right\} \right].P(f_n = c_i) \qquad (5)$$

where $p_{ji} = P(f_{n+1} = c_j \mid f_n = c_i)$

Here, f_{N+1} is interpreted as the end of the abstract (a virtual fragment after the last), and c_0 is its notional (fixed) component type.

$$p_{ij}est = \frac{seq(i,j)}{\min(N_i, 0.5)}$$

where *seq* is the same function as above, and in particular, *seq(i,*0*)* = number of occurrences of components of type *i* occurring at the end of abstracts.

Again, there are two cases:

Case (i): *n* = *N*

$$P^*(f_N = c_i) = p_{i0}.P(f_N = c_i) = \frac{seq(i,0)}{\min(N_i, 0.5)}.P(f_N = c_i)$$

Case (ii): *n* < *N*

$$P^*(f_n = c_i) = \left[\sum_{j=0}^{C} \left\{ seq(i,j).\frac{P^*(f_{n+1} = c_j)}{P(f_{n+1} = c_j)} \right\} \right].\frac{P(f_n = c_i)}{\min(N_i, 0.5)}$$

The two revision methods described above are not the same, and they may be used iteratively and/or alternately. In the description of experiments, below, structure-based revision strategies are denoted by strings of f's (for forward revision) and b's (for backward), *e.g.* fbfbfb.

4.4 Automatic text fragmentation

At the present time, automatic text fragmentation is very simple, and the problem cannot be regarded as adequately solved. Initially, the abstract is divided into fragments at punctuation characters. Two types of error can occur in this approximate analysis: the omission of a component boundary when it is not marked by punctuation, and the addition of a false component boundary when punctuation occurs within a component. In our data, the second type of error (extra false boundaries) is much more common than the first. Therefore, the idea of merging adjacent fragments under certain circumstances has been explored.

Two strategies have been devised for reducing the number of fragments in an attempt to approximate the known composition of components in abstracts:

(i) Merging adjacent fragments, according to their size (number of words) and the specific punctuation separating them.
(ii) Merging adjacent fragments, according to size and punctuation, but conditional upon there being a component type that has a comparatively high probability for both fragments.

Both procedures use a function, *w(p)*, of the punctuation, *p*, between fragments. *w(p)* is the proportion of occurrences of *p* that coincide with component boundaries in a sample corpus.

Strategy (i) ("merge small fragments") is applied before any probabilities are computed. Two adjacent fragments are merged into one if the size, in words, of either is less than or equal to *s*, where *s = a.w(p) + b* (*a* and *b* constants). The slope, *a*, of this linear function is negative, so the less likely the punctuation is to signal a component boundary, the longer the fragments that will be merged. In experiments, values of *a* and *b* yielding the following values of *s* have been used:

punctuation	*s*
,	4
;	4
:	4
?	3
.	1
!	1
other	2

Strategy (ii) can only be applied after component probabilities for the fragments of an abstract have been computed. In this strategy, two adjacent components, f_n and f_{n+1}, are merged if there exists a component type, c_j, for which $P(f_n = c_j) \times P(f_{n+1} = c_j) > H$, where

$$H = \frac{k.w(p)}{1 + 1/(size \cdot of \cdot f_{n+1})}$$

(0.75 has been found to be a good value for *k* when this merge step is applied before structural revision)

Probabilities are recalculated for merged fragments, using the combined sets of lexical clues.

5. Performance evaluation

How do we measure and compare performance of the discourse-level structure analysis programs? To some extent, this must depend upon the use to be made of the structures, and this is a little problematic because the IR system that will use them has not yet been designed. In general terms, there will be at least two uses:

(i) as an additional factor in matching or selection of documents;
(ii) as information to be included in the display of a retrieved abstract to the user.

Uncertainty in the structure may have a quite different impact on these two processes. For retrieval purposes, we could retain a number of alternative analyses, with associated probabilities, and design a suitable probabilistic matching function. For display, we may need to commit ourselves to the most likely analysis, although some vagueness could be allowed in the layout, and less certain aspects could be omitted.

In the following discussion, a "target" component is defined as a component type known to occur (through human text annotation) in the fragment.

5.1 Criteria for good performance

(i) Target components should be at the top of the list, ranked by probability estimate.
(ii) Target components should be close to the top of the ranked list.
(iii) Target components should have high probability.
(iv) Target components should have high probability relative to others.
(v) The fragmentation of the text should be close to that given by human experts.
(vi) Some errors are better than others. It is preferable that a fragment be mistaken for a component closely related in the hierarchy to the target, than for a more distant one.

Criterion (i) is more important for display purposes than for retrieval. Criteria (ii), (iii) and (iv) are related, but are not necessarily equivalent. It is possible to imagine on the one hand a fairly flat distribution of probabilities in which the target is more frequently near the top than it is, on the other hand, in a highly skewed distribution. Now, the probability distributions are constrained: the sum of the probabilities of all the component types for a particular text fragment is always 1. Therefore, the skew of the distribution is directly related to the magnitude of the highest probability. Throughout the experiments, it has been found that reasonably motivated versions of the processing model invariably produce mean target probabilities substantially higher than what one would expect from a flat distribution. In these circumstances, measuring the rank order of the target seems the most useful (criteria (i) and (ii)), and will capture criteria (iii) and (iv) quite well. (So far, criteria (v) and (vi) have not been systematically applied.)

5.2 Processing framework

The data available was generated by Elizabeth Liddy, in the course of her Ph.D dissertation research. [2] It is a collection of abstracts of empirical research papers and reports, obtained from the ERIC and PsycINFO databases. Each abstract has been segmented into its discourse-level components, according to Liddy's Elaborated model. This structuring was validated as part of the experimental work in her project, and is regarded as reliable. During subsequent work [1], each abstract was marked-up, manually, with a specially designed bracketing system, to facilitate computer processing of the structured abstracts. This corpus consists of the following:

eric.text	150 abstracts	1754 components
psyc.text	126 abstracts	1464 components

For present purposes a random sample of 75 abstracts was extracted from eric.text, forming eric.smpl. The complementary file is eric.cmpl:

eric.smpl	75 abstracts	915 components
eric.cmpl	75 abstracts	839 components

In the Elaborated model of the discourse-level structure, there are 37 component types, 15 of which are in the Typical model, and 7 of these are in the Prototypical model.

The first processing step is to extract clues, and various statistical information from some part of this corpus (usually eric.smpl).

The second stage is to analyze texts from one or more of the files, using the information from step one for estimation of probabilities, and concurrently to measure performance in relation to the components given by the bracketing.

In each step, there are several parameters that need to be set, so a very large number of runs are possible.

5.3 Performance measurement

The result of the analysis of each text fragment is a list of <component type, probability> pairs ranked in descending order of probability. We can thus obtain the probabilities and rank order of target components. Also calculated is a standardized score for each target: the ratio of the probability of the target to the probability at rank position 1 for the fragment. The mean probability and standardized score over the whole run is reported, as is the t-test score computed for the difference between probabilities actually obtained for target components and expected values for a uniform distribution.

The measure of overall quality of component ranking is calculated as follows (see Figure 4):

$$R_i = \left(\sum_{j=1}^{F} K_{ij} \right) \Big/ F \qquad \text{for } 1 \leq \text{i} \leq 10\text{, where F is the number of fragments analyzed in the run,}$$

and $K_{ij} = 1 / T_j$ where T_j = number of targets in fragment j if a target is ranked in position i, or
= 0, otherwise.

If the correspondence between fragment and component boundaries is exact, T_j is always 1, and R_i is just the proportion of targets ranked in position *i*. R_i are tabulated as percentages, along with cumulative percentages, beginning with rank 1. The normalized area under the cumulative curve is the final measure of performance, *M*. Specifically,

$$M = \left(\sum_{i=1}^{10} CR_i \right) \Big/ 10 \quad \text{where} \quad CR_i = \sum_{p=1}^{i} R_p$$

M can vary between 0 and 1. We would get 1 if the target were always ranked first, and fragmentation were always correct.

A method of testing the significance of differences between the *M*'s obtained from two different runs has yet to be decided. The problem is that if fragmentation differs from one run to another, it would not be possible to pair up the individual observations and do a matched-pairs test. On the other hand, an unrelated samples test seems weaker than necessary, because there would usually be substantial correspondences between the fragments. Hence, conclusions must be thoroughly hedged at this time.

6. Experiments and Results

6.1 Variables:

The discourse-level model described above is relatively simple. In other words, one can quite easily think of modifications which hold promise of improving the performance. Even within the confines of the present model, however, there are many possible variations. The variables which have been considered for investigation so far are as follows:

I. *Clue-set generation:* (see section 4.2)

I.1 sub-corpus used: eric.smpl, eric.cmpl or psyc.text (see section 5.2)
I.2 words and/or stems
I.3 lowfrequency filters:
A $[n_{ci} \leq A]$
B $[\sum_{c} n_{ci} \leq B]$
I.4 K = number of clues

II. *Fragmentation:* (see section 4.4)

"given" (*i.e.* using the manual marking-up of the test corpora) or "automatic" (*i.e.* using punctuation)

III. *Use of structural information:* (see section 4.3)

The strategy for dynamic revision of component probabilities, given by a string of f's (forward revision) and b's (backward revision)

IV. *Fragment merging,* for use with automatic fragmentation: (see section 4.4)

IV.1 "merge small fragments"
IV.1.1 fragment size parameters: a and b in $s = a.w(p) + b$

IV.2 "probabilistic merge"
IV.2.1 position of merge relative to structural revision strategy
IV.2.2 fragment matching parameter: k in formula for H

V. *Discourse structure model:* (see section 2)

Elaborated, Typical or Prototypical

VI. Sub-corpus analyzed:

Eric.smpl, eric.cmpl or psyc.text

6.2 Experimental runs:

All runs reported below relate to the Elaborated Model of the discourse-level structure of empirical abstracts, which has 37 component-types.

6.2.1 Clue-set runs

To find out how to generate clues, and establish a performance base line.

> [Early runs established low frequency filter values: *A* = 1 is best, *B* = 2 is good – little difference with *B* = 3.]
>
> These runs look at the number of clues, *K*, the use of stems, and sub-corpora variations.
>
> We generate *K* (varying from 0 to 1000) clues (words or words and stems) from eric.smpl, then apply lexical clues only (*i.e.* no structural revision) to eric.smpl, eric.cmpl and psyc.text, using given fragmentation into components. This looks at only the component probabilities estimation, assuming ideal text fragmentation. The results are given in Tables 1 – 4 and Figure 5.
>
> Comments:

1.1 When *K* = 0 (*i.e.* no clues), the model gives us the effect of using just relative frequency of occurrence of the various types of component. We can think of this as a benchmark.

1.2 For “words”, the maximum number of clues that can be extracted from the eric.smpl file (with *A* = 1, *B* = 2) is 670. With “words + stems”, this goes up to about 1100.

1.3 The shapes of the curves are similar – a general climb from the benchmarks.

1.4 Eric.smpl > eric.cmpl > psyc.text, as expected.

1.5 Benchmarks for eric.smpl and eric.cmpl are very close, reflecting the fact that the relative frequency distributions of component types in the two sub-corpora are very similar.

1.6 Adding stems does not do much for performance.

6.2.2 Structural information runs

> [Early runs established that performance increases as f → fb → fbfb → fbfbfb → fbfbfbfb. Improvement with the last step is very small, and run-time is increasing noticeably, so we use fbfbfb.]
>
> Clues are words from eric.smpl: *A* = 1, *B* = 2, *K* = {0, 120, 670}. The given fragmentation into components is used. Sub-corpora analyzed are eric.smpl, eric.cmpl, and psyc.text. The results are given in Tables 5 – 7 and Figure 6.
>
> Comments:

2.1 Good improvements are observed for all values of *K* and each sub-corpus.

2.2 Differences for *K* = 0 indicates that structure-based probability revision is beneficial even in the absence of lexical clues.

2.3 Difference in the performance at *K* = 0 for eric.smpl and eric.cmpl indicates differences in the adjacency profiles between the two sub-corpora.

6.2.3 Automatic fragmentation runs

These runs test the use of punctuation characters to divide text into fragments.

Clues are words or words + stems from eric.smpl: *A* = 1, *B* = 2, *K* from 0 to 1000. The sub-corpora analyzed are eric.smpl, eric.cmpl, and psyc.text. The results are given in Table 8 – 11 and Figure 7.

Comments:

3.1 Extraordinary! Performance declines as more clues are added (Benchmark > lexical clues).

3.2 Pattern is repeated in all three sub-corpora.

3.3 Eric.smpl > eric.cmpl > psyc.text (analyzed using estimates from eric.smpl) as expected.

3.4 Adding stems adds little to performance.

3.5 Why the degradation?

a) If a component is divided between two or more fragments, its clues will also be divided and this will have two effects: (i) clues for one component will not reinforce each other; and (ii) some parts of the component may have no clues at all.

b) If a fragment contains parts of more than one component, the clues may work against each other.

6.2.4 Automatic fragmentation with structural information

These runs test the use of punctuation characters to do initial fragmentation, followed by structural revision strategy S = fbfbfb and/or fragment merging.

Clues are words from eric.smpl: *A* = 1, *B* = 2, *K* = {0, 120, 670}. Fragment merging methods are “small fragments” (m_s), and “probabilistic” (m_p) with *k* = 0.75. The combinations tested are: m_s, m_sm_p, S, m_sS, m_sm_pS. The sub-corpora analyzed are: eric.smpl, eric.cmpl, and psyc.text. Results are shown in Tables 12 – 14 and Figures 8 – 10.

Comments:

4.1 Use of structural information has a large impact, raising performance above benchmark levels.

4.2 Merging small fragments has a small beneficial effect.
4.3 Merging "probabilistically" has a larger effect.
4.4 The effects do compound.
4.5 The patterns for analyzing eric.cmpl and psyc.text (using estimates from eric.smpl) are similar to that for eric.smpl, though less pronounced.

Conclusions:

The results are encouraging enough to indicate that further work would be worthwhile. The most important results are those obtained from runs in which clues and text-structure distributions were obtained from one sub-corpus (eric.smpl) and then applied to the other sub-corpus (eric.cmpl). This approximates a real-life application in which a sample of texts of a particular type are analyzed by hand to provide statistical information that can be used with a large collection of similar texts. When the fragmentation into components is given, the M value reaches 0.722 (Table 6). With the best automatic fragmentation methods devised so far, the M value reaches 0.695 (Table 13). Applying clues obtained from texts on education to the task of discourse-level analysis of texts on psychology clearly does not work so well, presumably because the vocabularies are different. Some limitations of this study are: (i) the corpus used (and hence the sub-corpus used for clue-derivation) is very small; (ii) the types of clues were limited to words and/or stems; (iii) the basic automatic fragmentation technique is based only upon punctuation. Further development in any of these areas can be expected to improve what are already quite good results. An additional, important limitation of the work is that the only text type included is the empirical abstract in the behavioral science literature. Other text-types should be explored.

References:

1. Oddy, R.N., Liddy, E.D., Balakrishnan, B., Bishop, A., Elewononi, J., Martin, E. "Towards the Use of Situational Information in Information Retrieval." *Journal of Documentation, 48*(2), 1992, 123-171.
2. Liddy, E.D. "The Discourse-level Structure of Empirical Abstracts: an Exploratory Study." *Information Processing and Management, 27*(1), 1991, 55-81.

clues	mean target probability	% targets at rank 1	% targets up to rank 10	M
0	0.063	11	69	0.462
20	0.139	24	79	0.572
50	0.188	31	81	0.620
80	0.221	34	84	0.648
120	0.270	39	84	0.673
160	0.296	41	84	0.683
204	0.312	42	84	0.680
302	0.339	44	85	0.677
400	0.424	50	86	0.708
500	0.485	52	86	0.728
600	0.521	55	85	0.738
670	0.564	58	86	0.766

Table 1: Model: Elaborated
Fragments: given
Clues: words from eric.smpl
Applied to: eric.smpl

clues	mean target probability	% targets at rank 1	% targets up to rank 10	M
0	0.063	11	69	0.462
20	0.143	22	77	0.556
60	0.209	30	81	0.607
83	0.250	33	83	0.636
120	0.299	40	85	0.675
163	0.330	42	87	0.691
205	0.349	43	87	0.696
300	0.395	46	87	0.706
408	0.417	48	86	0.708
500	0.454	51	87	0.717
600	0.488	54	87	0.730
700	0.539	57	87	0.756
800	0.568	59	87	0.761
900	0.583	61	86	0.764
1000	0.611	63	86	0.785

Table 2: Model: Elaborated
Fragments: given
Clues: words and stems from eric.smpl
Applied to: eric.smpl

clues	**mean target probability**	**% targets at rank 1**	**% targets up to rank 10**	**M**
0	0.062	9	68	0.460
20	0.115	18	74	0.526
50	0.154	24	75	0.558
80	0.179	26	77	0.580
120	0.213	29	76	0.597
160	0.239	31	77	0.603
204	0.244	29	76	0.594
302	0.255	33	78	0.595
400	0.305	37	79	0.627
500	0.362	41	80	0.644
600	0.371	41	78	0.640
670	0.404	42	80	0.667

Table 3: Model: Elaborated
Fragments: given
Clues: words from eric.smpl
Applied to: eric.cmpl

clues	**mean target probability**	**% targets at rank 1**	**% targets up to rank 10**	**M**
0	0.057	11	60	0.414
20	0.083	16	61	0.443
50	0.109	21	62	0.465
80	0.131	23	63	0.465
120	0.153	24	62	0.480
160	0.165	25	62	0.479
204	0.174	24	62	0.480
302	0.185	23	61	0.478
400	0.231	27	63	0.497
500	0.256	29	62	0.497
600	0.268	29	62	0.501
670	0.302	32	63	0.520

Table 4: Model: Elaborated
Fragments: given
Clues: words from eric.smpl
Applied to: psyc.text

clues	mean target probability	% targets at rank 1	% targets up to rank 10	M
0	0.135	19	83	0.578
120	0.342	45	89	0.720
670	0.604	62	90	0.806

Table 5: Model: Elaborated
Fragments: given
Clues: words from eric.smpl
Structure revision: fbfbfb
Applied to: eric.smpl

clues	mean target probability	% targets at rank 1	% targets up to rank 10	M
0	0.124	15	79	0.552
120	0.283	35	84	0.660
670	0.444	46	85	0.722

Table 6: Model: Elaborated
Fragments: given
Clues: words from eric.smpl
Structure revision: fbfbfb
Applied to: eric.cmpl

clues	mean target probability	% targets at rank 1	% targets up to rank 10	M
0	0.091	13	74	0.493
120	0.210	27	76	0.566
670	0.337	37	75	0.589

Table 7: Model: Elaborated
Fragments: given
Clues: words from eric.smpl
Structure revision: fbfbfb
Applied to: psyc.text

clues	mean target probability	% targets at rank 1	% targets up to rank 10	M
0	0.070	32	74	0.562
50	0.123	20	77	0.525
120	0.153	22	80	0.509
204	0.163	22	79	0.506
302	0.170	22	70	0.479
500	0.213	23	68	0.489
670	0.231	24	65	0.489

Table 8: Model: Elaborated
Fragments: automatic
Clues: words from eric.smpl
Applied to: eric.smpl

clues	mean target probability	% targets at rank 1	% targets up to rank 10	M
0	0.070	32	74	0.562
60	0.130	18	77	0.514
120	0.165	24	81	0.508
205	0.180	23	82	0.516
300	0.194	24	82	0.519
500	0.208	23	74	0.494
700	0.231	25	68	0.497
1000	0.254	26	66	0.503

Table 9: Model: Elaborated
Fragments: automatic
Clues: words and stems from eric.smpl
Applied to: eric.smpl

clues	**mean target probability**	**% targets at rank 1**	**% targets up to rank 10**	**M**
0	0.070	31	76	0.567
50	0.107	16	74	0.492
120	0.125	20	76	0.468
204	0.129	17	75	0.457
302	0.125	14	62	0.415
500	0.148	15	63	0.432
670	0.149	16	59	0.401

Table 10: Model: Elaborated
Fragments: automatic
Clues: words from eric.smpl
Applied to: eric.cmpl

clues	**mean target probability**	**% targets at rank 1**	**% targets up to rank 10**	**M**
0	0.062	32	63	0.486
50	0.078	13	60	0.392
120	0.104	16	59	0.364
204	0.108	14	59	0.357
302	0.102	12	47	0.324
500	0.119	13	47	0.326
670	0.111	12	44	0.300

Table 11: Model: Elaborated
Fragments: automatic
Clues: words from eric.smpl
Applied to: psyc.text

merge/ structure	**clues**	**fragment ratio**	**mean target probability**	**% targets at rank 1**	**% targets up to rank 10**	**M**
$m_s m_p S$	0	1.45	0.331	33	76	0.553
	120	1.29	0.359	36	86	0.685
	670	0.88	0.490	50	90	0.778
$m_s S$	0	1.45	0.331	33	76	0.553
	120	1.45	0.352	35	86	0.667
	670	1.45	0.423	44	88	0.735
S	0	1.56	0.330	33	75	0.550
	120	1.56	0.350	35	86	0.663
	670	1.56	0.408	43	88	0.726
$m_s m_p$	0	1.45	0.070	32	75	0.562
	120	1.29	0.163	23	80	0.512
	670	0.88	0.362	38	78	0.641
m_s	0	1.45	0.070	32	75	0.562
	120	1.45	0.160	24	80	0.519
	670	1.45	0.245	25	66	0.505

Notes: merge/structure m_s = merge small fragments
m_p = probabilistic merge
S = structure revision strategy fbfbfb
fragment ratio = number of fragments determined automatically / number given

Table 12: Use of Structural Information and Fragment Merging
Model: Elaborated
Fragments: automatic
Clues: words from eric.smpl
Applied to: eric.smpl

merge/ structure	clues	fragment ratio	mean target probability	% targets at rank 1	% targets up to rank 10	M
m_sm_pS	0	1.52	0.316	31	74	0.536
	120	1.31	0.343	34	84	0.659
	670	0.89	0.394	40	84	0.695
m_sS	0	1.52	0.316	31	74	0.536
	120	1.52	0.336	34	84	0.648
	670	1.52	0.373	39	84	0.675
S	0	1.63	0.315	31	74	0.531
	120	1.63	0.335	34	84	0.646
	670	1.63	0.359	38	85	0.673
m_sm_p	0	1.52	0.070	31	76	0.563
	120	1.31	0.135	20	75	0.464
	670	0.89	0.220	23	69	0.503
m_s	0	1.52	0.070	31	76	0.563
	120	1.52	0.128	21	75	0.471
	670	1.52	0.158	17	60	0.410

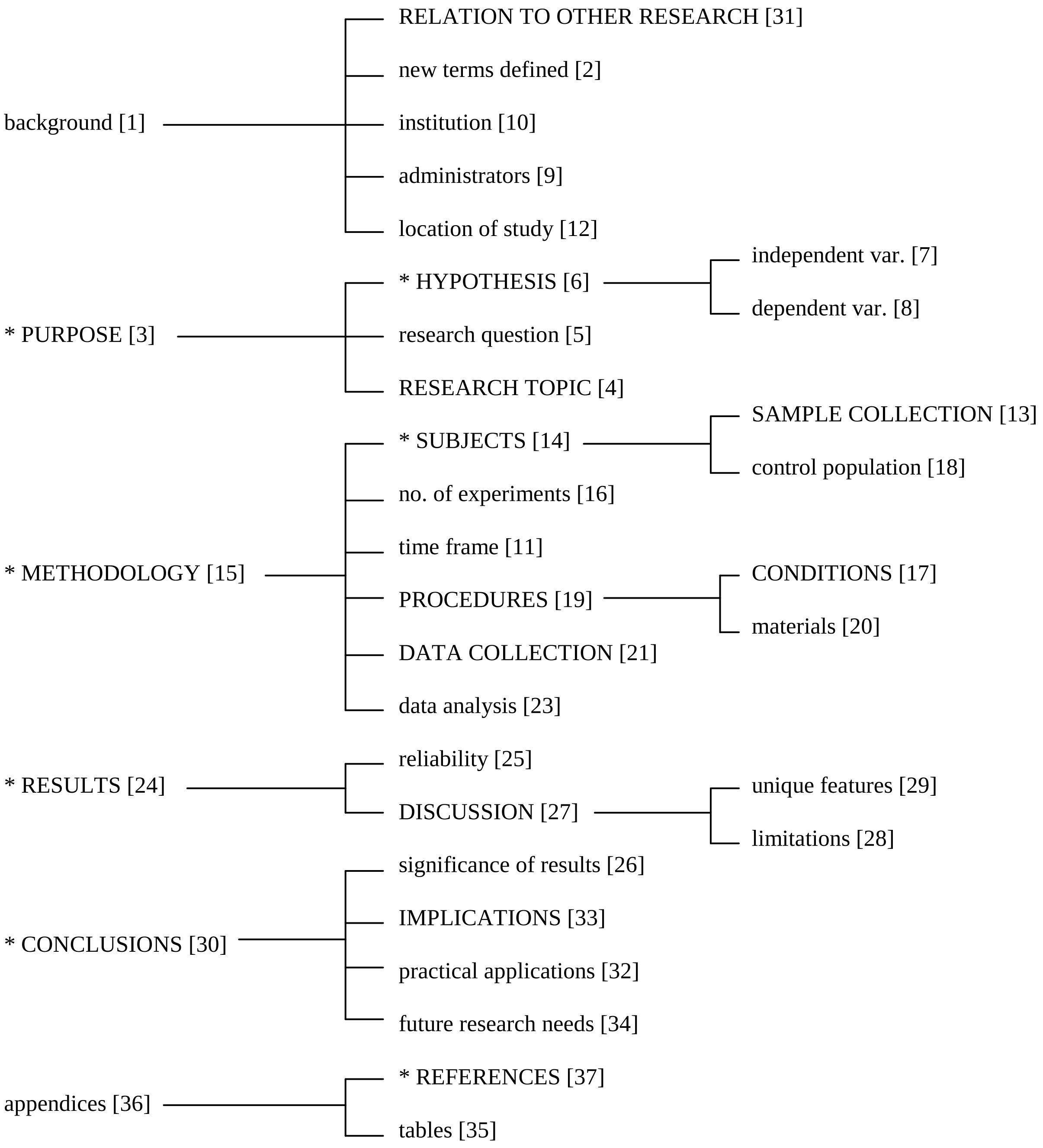
Notes: merge/structure m_s = merge small fragments
m_p = probabilistic merge
S = structure revision strategy fbfbfb
fragment ratio = number of fragments determined automatically / number given

Table 13: Use of Structural Information and Fragment Merging
Model: Elaborated
Fragments: automatic
Clues: words from eric.smpl
Applied to: eric.cmpl

merge/ structure	clues	fragment ratio	mean target probability	% targets at rank 1	% targets up to rank 10	M
m_sm_pS	0	1.26	0.329	33	69	0.498
	120	1.17	0.335	34	77	0.570
	670	0.80	0.349	35	77	0.603
m_sS	0	1.26	0.329	33	69	0.498
	120	1.26	0.337	34	77	0.568
	670	1.26	0.339	36	73	0.579
S	0	1.34	0.323	32	67	0.483
	120	1.34	0.330	33	77	0.557
	670	1.34	0.329	35	72	0.567
m_sm_p	0	1.26	0.063	32	64	0.493
	120	1.17	0.105	16	59	0.366
	670	0.80	0.165	18	55	0.396
m_s	0	1.26	0.063	32	64	0.493
	120	1.26	0.108	17	60	0.372
	670	1.26	0.114	12	46	0.310

Notes: merge/structure m_s = merge small fragments
m_p = probabilistic merge
S = structure revision strategy fbfbfb
fragment ratio = number of fragments determined automatically / number given

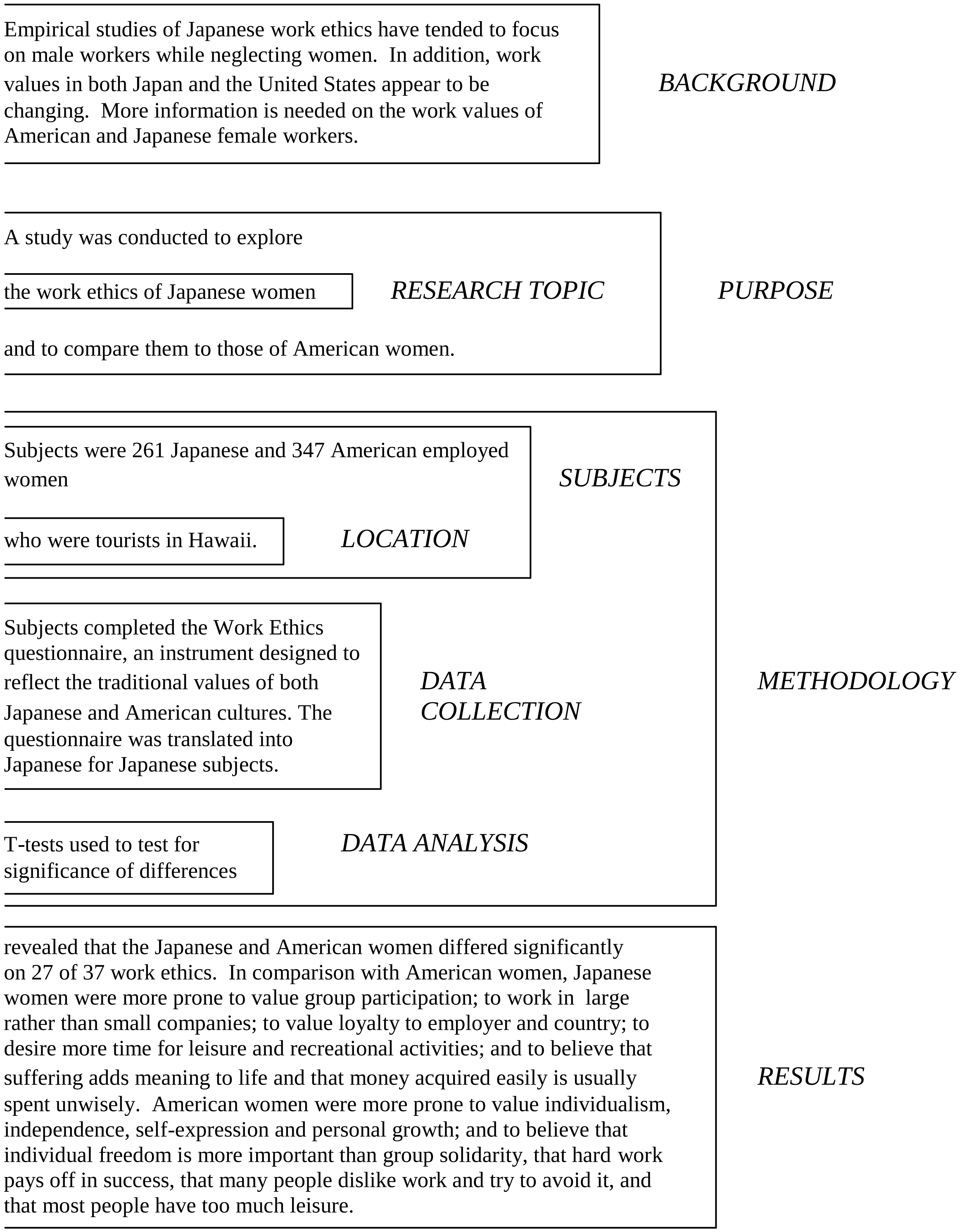
Table 14: Use of Structural Information and Fragment Merging
Model: Elaborated
Fragments: automatic
Clues: words from eric.smpl
Applied to: psyc.text



(* - Prototypical component; UPPER-CASE lettering – Typical component; all – elaborated components)

Figure 1: Structure of Empirical Abstracts

ER15



Empirical studies of Japanese work ethics have tended to focus on male workers while neglecting women. In addition, work values in both Japan and the United States appear to be changing. More information is needed on the work values of American and Japanese female workers.

A study was conducted to explore

the work ethics of Japanese women

and to compare them to those of American women.

Subjects were 261 Japanese and 347 American employed women

who were tourists in Hawaii.

Subjects completed the Work Ethics questionnaire, an instrument designed to reflect the traditional values of both Japanese and American cultures. The questionnaire was translated into Japanese for Japanese subjects.

T-tests used to test for significance of differences

revealed that the Japanese and American women differed significantly on 27 of 37 work ethics. In comparison with American women, Japanese women were more prone to value group participation; to work in large rather than small companies; to value loyalty to employer and country; to desire more time for leisure and recreational activities; and to believe that suffering adds meaning to life and that money acquired easily is usually spent unwisely. American women were more prone to value individualism, independence, self-expression and personal growth; and to believe that individual freedom is more important than group solidarity, that hard work pays off in success, that many people dislike work and try to avoid it, and that most people have too much leisure.

Figure 2: A Structured Abstract

```
ER15
[1+ Empirical studies of Japanese work ethics have tended to focus
on male workers while neglecting women.  In addition, work values
in both Japan and the United States appear to be changing.  More
information is needed on the work values of American and Japanese
female workers. 1]
[3+ A study was conducted to explore
[4+ the work ethics of Japanese women 4]
and to compare them to those of American women. 3]
[14+Subjects were 261 Japanese and 347 American employed women
[12+who were tourists in Hawaii. 12] 14]
[21+ Subjects completed the Work Ethics questionnaire, an
instrument designed to reflect the traditional values of both
Japanese and American cultures. The questionnaire was translated
into Japanese for Japanese subjects. 21]
[24+ [23+ T-tests used to test for significance of differences 23]
revealed that the Japanese and American women differed
significantly on 27 of 37 work ethics.  In comparison with
American women, Japanese women were more prone to value group
participation; to work in large rather than small companies; to
value loyalty to employer and country; to desire more time for
leisure and recreational activities; and to believe that suffering
adds meaning to life and that money acquired easily is usually
spent unwisely.  American women were more prone to value
individualism, independence, self-expression and personal growth;
and to believe that individual freedom is more important than
group solidarity, that hard work pays off in success, that many
people dislike work and try to avoid it, and that most people have
too much leisure. 24]
```

Figure 3. An Abstract, marked up with component brackets

FRAGMENT:

```
[24+ [23+ T-tests used to test for significance
of differences 23] revealed that the Japanese
and American women differed significantly on
27 of 37 work ethics.
```

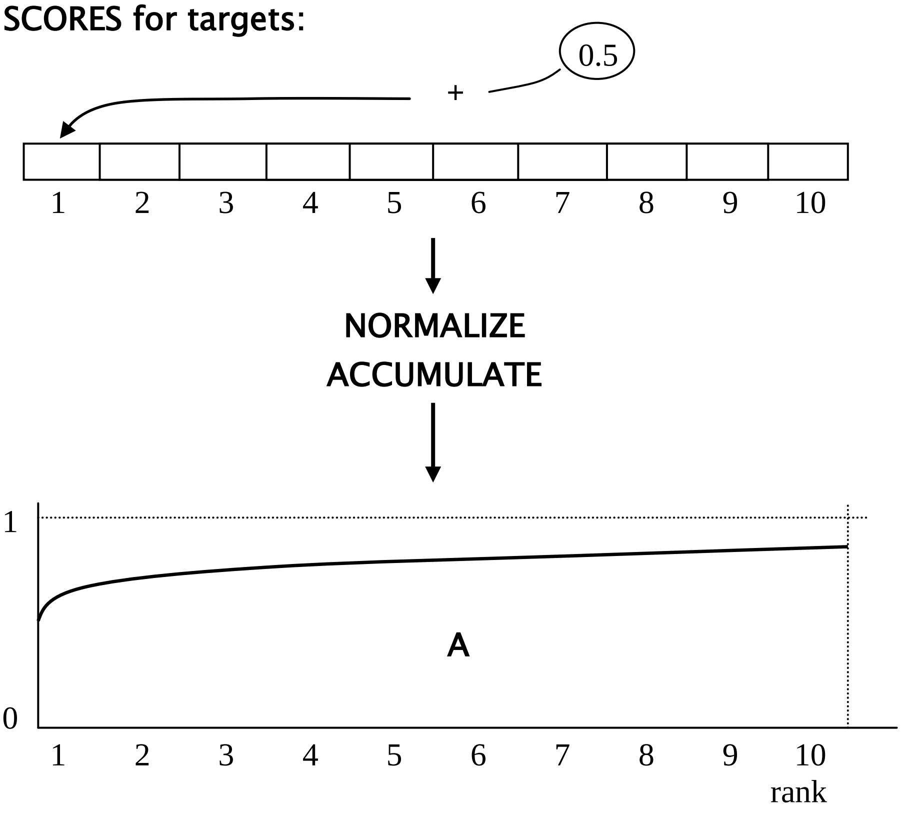
TARGETS: 23, 24

CLUES: differ, differed, differences, reveal, revealed, signif, significantly, t, that

PROBABILITIES:

1	24	0.993895
2		
.		
17	23	0.85×10^{-6}
.		
37		
mean for targets:		0.49699

RANK SCORES for targets:



Evaluation measure, M = area A / 10

Figure 4: Component Ranking Evaluation Procedure

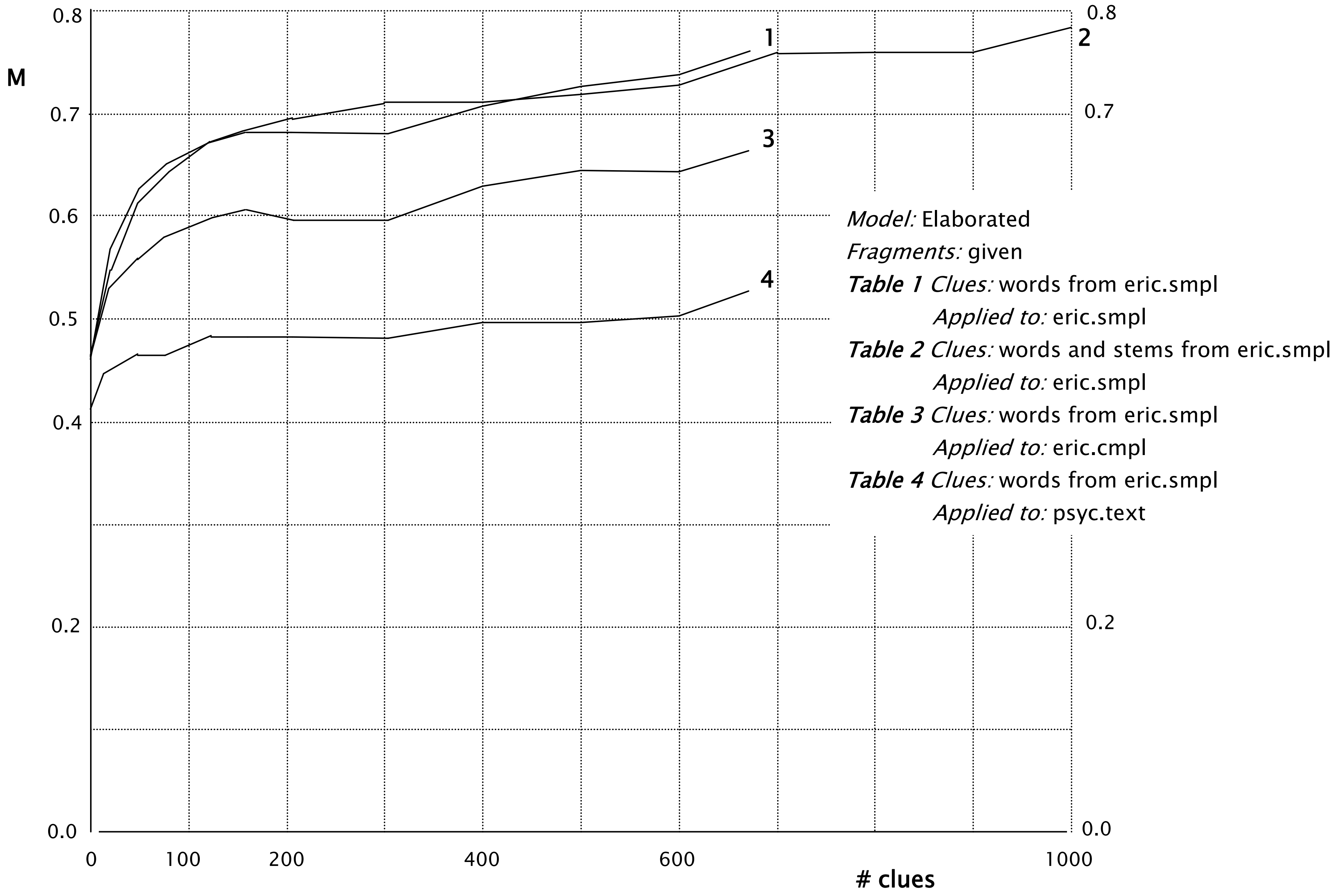


Figure 5: Component identification from lexical clues

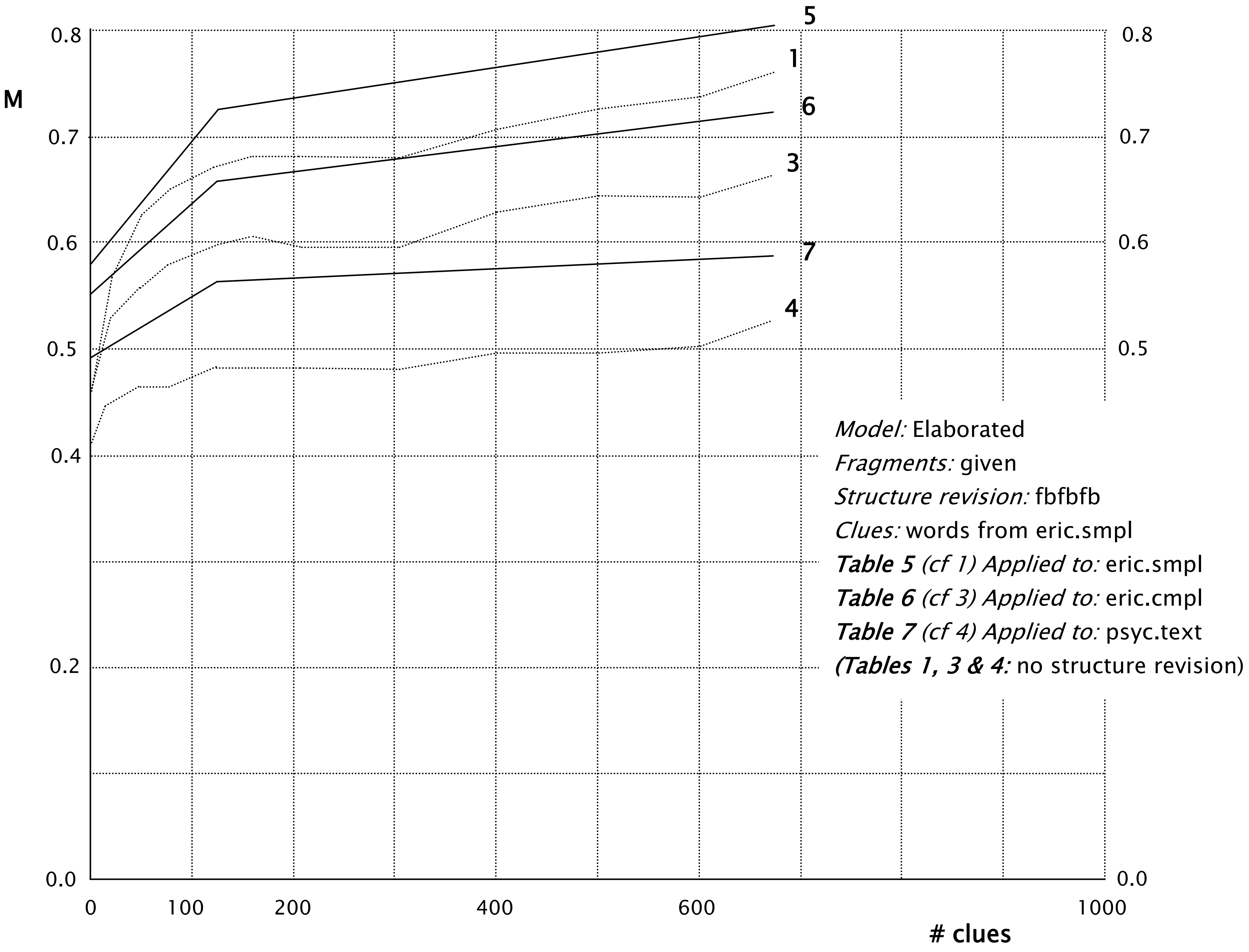


Figure 6: Component identification from lexical clues and structural information

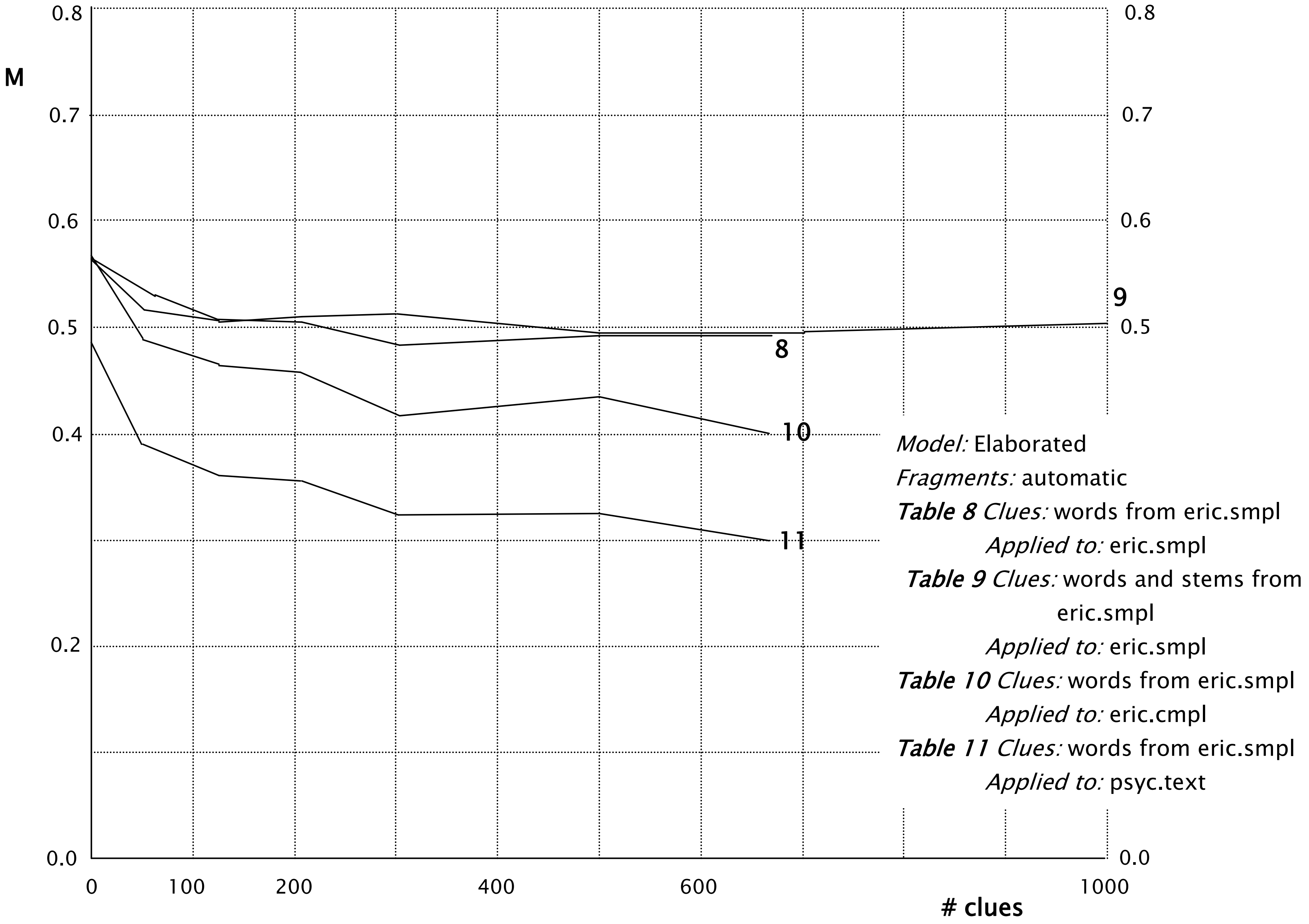


Figure 7: Component identification with automatic fragmentation, from lexical clues

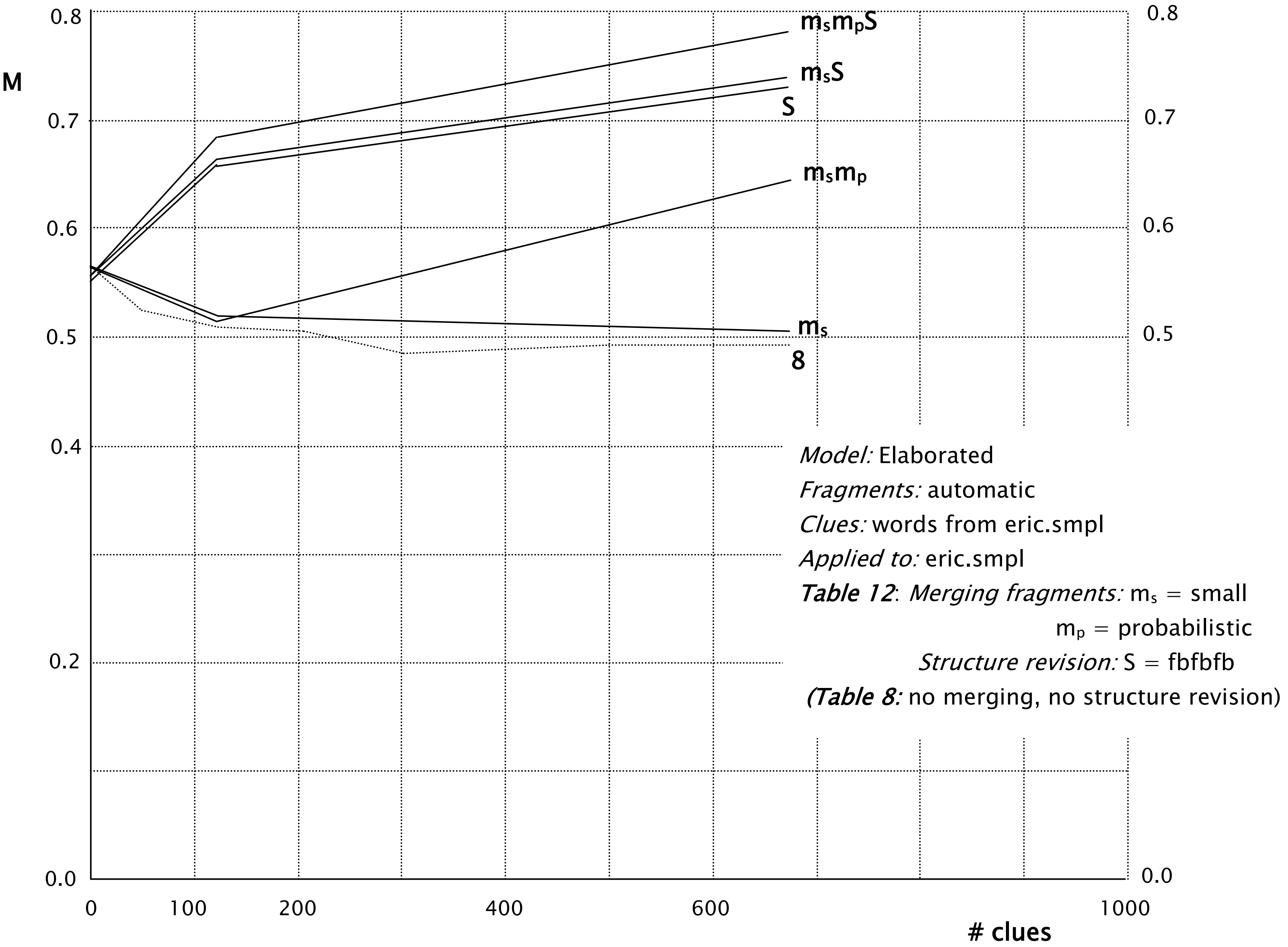


Figure 8: Component identification with automatic fragmentation and merging, from lexical clues and structural information
Clues: words from eric.smpl   Applied to: eric.smpl

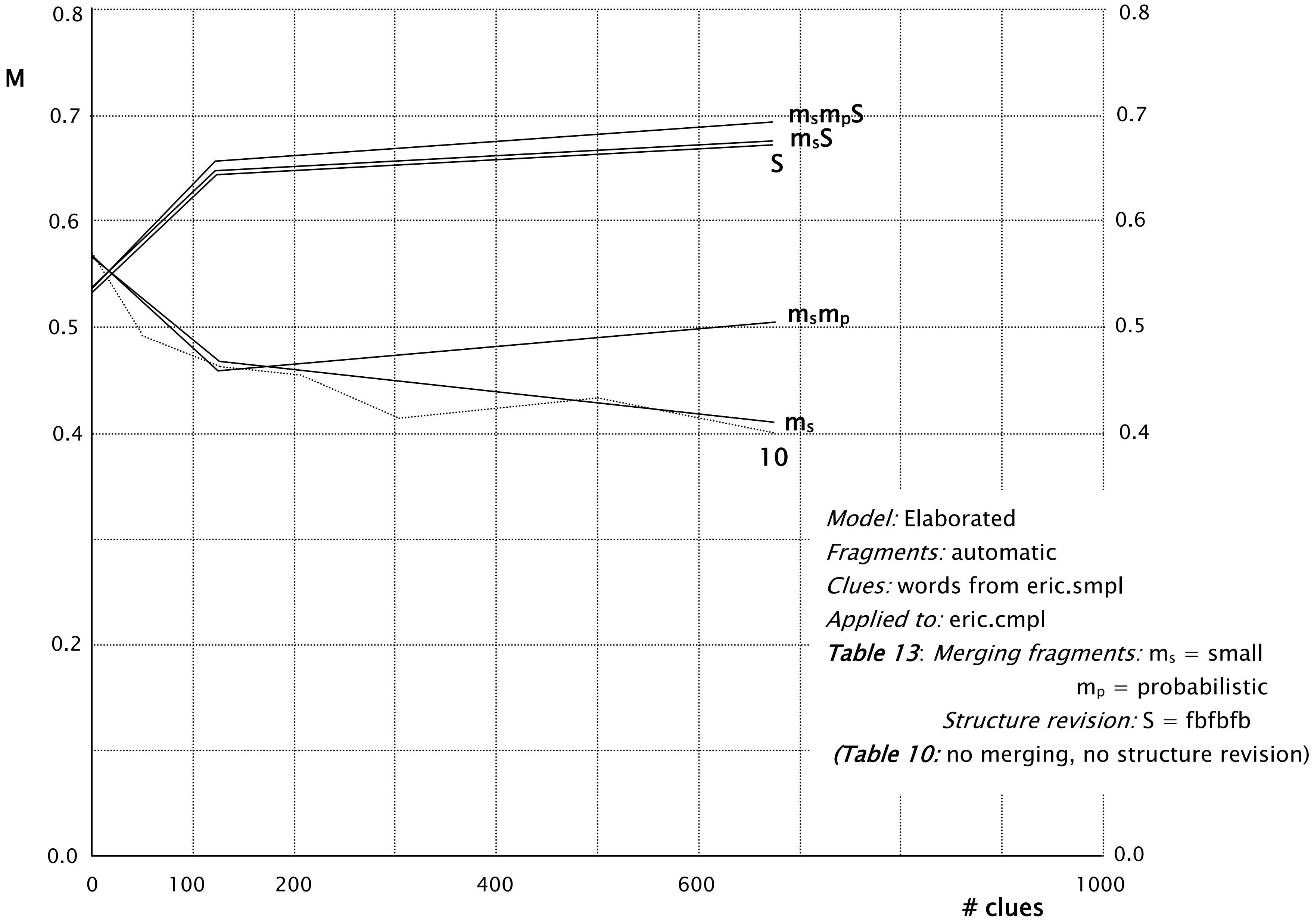
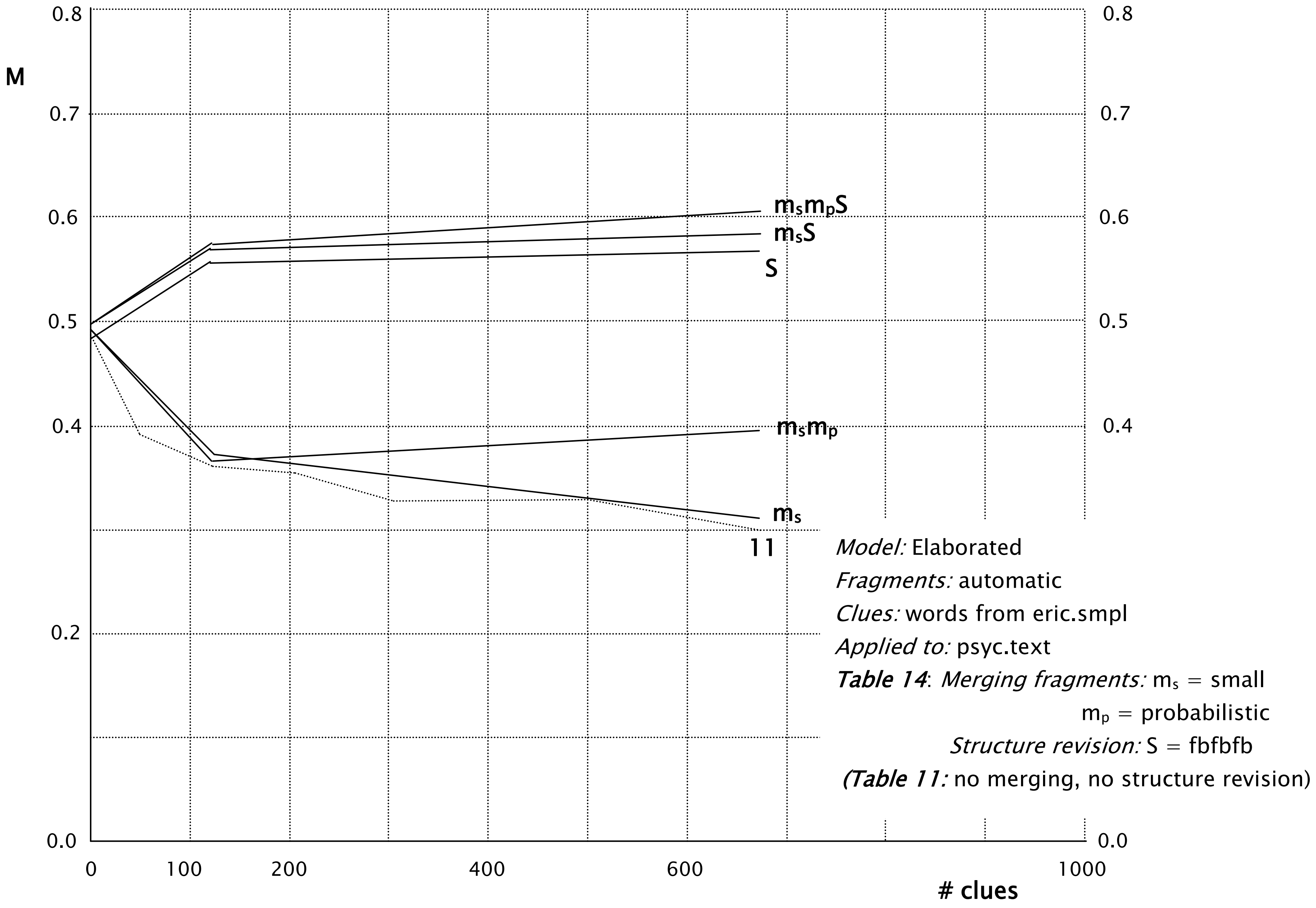


Figure 9: Component identification with automatic fragmentation and merging, from lexical clues and structural information
Clues: words from eric.smpl Applied to: eric.cmpl



Model: Elaborated

Fragments: automatic

Clues: words from eric.smpl

Applied to: psyc.text

Table 14: *Merging fragments:* m_s = small
m_p = probabilistic

Structure revision: S = fbfbfb

(***Table 11:*** no merging, no structure revision)

Figure 10: Component identification with automatic fragmentation and merging, from lexical clues and structural information

Clues: words from eric.smpl Applied to: psyc.text