# CATCHING THE CHESHIRE CAT

Christer Johansson

Dept. of Linguistics, Lund University, Sweden
email: Christer.Johansson@ling.lu.se

**ABSTRACT**
Finding useful phrases is important in applications like information retrieval, and text-to-speech systems. One of the currently most used statistics is the mutual information ratio. This paper compares the mutual information ratio and a measure that takes temporal ordering into account. Using this modified measure, some local syntactic constraints as well as phrases are captured.

## INTRODUCTION

In *Alice's Adventures in Wonderland* by Lewis Carrol many of Alice's friends have names that consists of two words, for example: the March Hare, the Mock Turtle, and the Cheshire Cat. The individual words in these combinations, if we ignore capitalisation, might be quite common.

Individual words usually mean different things when they are free. For example, in "The March against Apartheid", and "The March Hare", "march" means totally different things. There is obviously a strong link between "the" and "march", but the link between "march" and "hare" is definitely stronger, at least in Lewis Carrol's text.

The goal of this paper is to propose a statistic that measures the strength of such glue between words in a sampled text. Finding the names of Alice's friends can be done by searching for two adjacent words with initial capital letters.

One use of statistical associations could be to find translatable concepts and phrases, that might be expressed with a different number of words in another language. Another possibly interesting use of statistical associations is to predict whether words constitute new or given information in speech. It has been proposed (e.g. Horne & Johansson, 1993) that the stress of words in speech is highly dependent on the informational content of the word. Also, statistical associations are not incompatible with the first stages of the "hypothesis space" proposed by Processability Theory (personal communication with Manfred Pienemann of Sydney University, see also Meisel & al., 1981).

There are different methods of calculating statistical associations. Yang & Chute (1992) showed that a linear least square mapping of natural language to canonical terms is both feasible, and a way of detecting synonyms. Their method does not seem to detect dependencies in the order of words however. To do this we need a measure that is sensitive to the order between words. In this paper we will use a variant of mutual information that derives from Shannon's theory of information. (as discussed in e.g., Salton & McGill, 1983)

### Definitions and assumptions

The definition of a word in a meaningful way is far from easy, but a *working definition*, for technical purposes, is to assume that a word equals a string of letters. These 'words' are separated by non-letters. The case of letters is ignored, i.e. converted into lower case. For example: "there's" are two 'words': "there" and "s".

A *collocation* consists of a word and the word that *immediately* follows. Index 1 will refer to the first word and 2 to the second word. Index 12 will refer to word 1 followed by word2, and similarly for 21.

Another assumption is that natural language is more predictive in the (left-to-right) temporal order, than in the reversed order. This is motivated by the simple observation that speech comes into the system through the ears serially.

For example: consider the French phrase "un bon vin blanc" (Lit. "a good wine white"). "Bon" can (relatively often) be followed by "vin", but usually not "vin" by "bon". The same kind of link exists between "vin" and "blanc", but not between "blanc" and "vin". This linking affects the intonation of French phrases, and *also* that intonation supports these kinds of links. Note, that this is not an explanation of either intonation or syntax: we most likely have to consider massive interaction between different modalities of language.

**Deriving the measure**

*The mutual information ratio,* μ, provides a rough estimation on the glue between words. It measures, roughly, how much more common a collocation is in a text than can be accounted for by chance. This measure does not assume any ordering between the words making up a collocation, in the sense that the μ-measure of [w1...w2] and [w2...w1] are calculated as if they were unrelated collocations.

The mutual information ratio (in Steier & Belew, 1991) is expressed:

$$\mu = \log_2\left(\frac{p([w_1 \cdots w_2])}{p(w_1)p(w_2)}\right)$$

**Formula 1**: The mutual information ratio

where 'p' defines the probability function, p([w1...w2]) is read as "the probability of finding word $w_2$ after word $w_1$".

**Adjusting for order between words**

We have experimented with the difference in mutual information, Δμ, between the two different orderings of two words making up a collocation. The results indicate that Δμ captures some of the local constraints in a sampled text. Δμ can be expressed:

$$\Delta\mu = \log_2\left(\frac{p([w1 \cdots w2])}{p(w1)p(w2)}\right) - \log_2\left(\frac{p([w2 \cdots w1])}{p(w1)p(w2)}\right)$$

=>

$$\log_2\left(\frac{p([w1 \cdots w2])}{p([w2 \cdots w1])}\right) = \log_2\left(\frac{F([w1 \cdots w2])}{F([w2 \cdots w1])}\right)$$

**Formula 2:** The difference in mutual information

where $F([w_X \ldots w_Y])$ denotes the frequency of which $w_X$ and $w_Y$ co-occur in the sample. $F(w_X)$ is the frequency of word $w_X$. Note that the size of the sample cancels in this equation. Note also that this measure is not sensitive to the individual probabilities of the words.

A problem is when there is no $F([w_2 \ldots w_1])$. In these cases, we have chosen to arbitrarily set $F([w_2 \ldots w_1])$ to 0.1, with the justification that *if* the sample was ten times larger we might have found at least one such pair.

**MATERIAL**

The material is *Alice's Adventures in Wonderland* by Lewis Carrol, available in electronic format via email from the Gutenberg Project. The text contains 27332 words of which 2576 are unique, making up a total of 14509 unique word pairs. Alice in Wonderland was chosen because it is a well-known text, it contains some phrases that we know are in there (e.g. *March Hare*), and it contains a sufficient number of words, and variations of words, to be interesting for the experiment. Studies could be done for other collections of texts, e.g. medical abstracts. As more documents are available, comparisons between documents can be done (Steier & Belew, 1991). This experiment only contains within comparisons of phrases for one specific text.

**METHOD**

For each of the unique words in the text the frequencies of all immediately following words were collected. In this text, no filtering of the text was performed. Some initial experiments were performed, with a stoplist, to remove function words and some other common words (see Fox, 1992, for details). Some simple stemming was also tried, e.g. removing 's' and 'ed' from the end of words. Stemming may lead to difficulties in distinguishing compounds from noun-verb complexes. It is not clear if the pros of using stemming outweighs the cons, consequently we decided to work with the raw text. Stoplists and stemming might be more important when the ordinary μ-measure is used.

**RESULTS**

The collocations were ordered differently by the two measures. The μ was sensitive to individual frequencies, and favoured very low frequency collocations. The Δμ was sensitive to the ordering of the words, and favoured high frequency collocations that only occurred in one order. The quality of the different measures can be seen by comparing the top and last ten collocations between the measures. **Table** 1.1 and 2.1 refer to Δμ, and **Table** 1.2 and 2.2 refer to μ. The N column tells the rank-number of the collocation. Note that the frequencies of the individual words, F1 and F2, are not used to compute Δμ, they are only provided for comparison with the μ-measure.

Note that the numerical values of the μ-measure and the Δμ-measure cannot be directly compared since they measure slightly different phenomena.

**Table 1.1:** The top ten collocations by Δμ

| N | word pair | Δμ | $F_1$ | $F_2$ | $F_{12}$ | $F_{21}$ |
|---|---|---|---|---|---|---|
| 1 | said-> the | 11.0 | 462 | 1642 | *210* | *0* |
| 2 | of-> the | 10.4 | 513 | 1642 | *132* | *0* |
| 3 | in->a | 9.92 | 369 | 632 | *97* | *0* |
| 4 | and-> the | 9.70 | 872 | 1642 | *83* | *0* |
| 5 | in-> the | 9.64 | 369 | 1642 | *80* | *0* |
| 6 | to-> the | 9.43 | 729 | 1642 | *69* | *0* |
| 7 | don->t | 9.25 | 61 | 218 | *61* | *0* |
| 8 | as-> she | 9.25 | 263 | 552 | *61* | *0* |
| 9 | a-> little | 9.20 | 632 | 128 | *59* | *0* |
| 10 | she-> had | 9.20 | 552 | 178 | *59* | *0* |

Δμ gives a measure of *local* links between words. As can be seen from **Table** 1.1, Δμ captures local constraints: that prepositions are usually followed by a noun phrase, that 'and' usually is used as a noun co-ordinator (indicated by the high value for 'and->the'). Mitjushin (1992) has proposed similar links on a higher syntactic level, using a rule-based approach. We have deliberately tried to avoid talking about word-classes since it is misleading at this level of analysis. However, we get many examples of good representatives for word-classes that form collocations.

**Table 1.2:** The top ten collocations by μ

| N | word pair | μ | $F_1$ | $F_2$ | $F_{12}$ |
|---|---|---|---|---|---|
| 1 | wooden->spades | 14.7 | *1* | *1* | *1* |
| 2 | various->pretexts | 14.7 | *1* | *1* | *1* |
| 3 | uncommonly->fat | 14.7 | *1* | *1* | *1* |
| 4 | turkey->toffee | 14.7 | *1* | *1* | *1* |
| 5 | tittered->audibly | 14.7 | *1* | *1* | *1* |
| 6 | tinkling->sheep | 14.7 | *1* | *1* | *1* |
| 7 | tide-> rises | 14.7 | *1* | *1* | *1* |
| 8 | tart->custard | 14.7 | *1* | *1* | *1* |
| 9 | steam->engine | 14.7 | *1* | *1* | *1* |
| 10 | splendidly->dressed | 14.7 | *1* | *1* | *1* |

The flavour of the collocations that μ rate highly is different. As can be seen from **Table** 1.2, low individual frequencies result in a high μ-value, even if the collocation is unique. This gives an illusion of a semantic relation, which is due to the fact that low frequency words are usually high in content. The μ-measure is useful when we are interested in the correlation between words within and between documents (Steier & Belew, 1991). This notion could be expanded upon to incorporate correlation between any two words in general, and it seems to work well for the μ-measure (Wettler and Rapp, 1989).

*The last ten collocations.* Δμ is sensitive to deviation from an expected ordering in the sample. The negative valued link between these words makes a phrase boundary between the two words probable.

**Table 2.1:** The last ten collocations by Δμ

| N | word pair | Δμ | $F_1$ | $F_2$ | $F_{12}$ | $F_{21}$ |
|---|---|---|---|---|---|---|
| 14500 | caterpillar-> the | -4.70 | 28 | 1642 | *1* | *26* |
| 14501 | mouse->the | -4.81 | 44 | 1642 | *1* | *28* |
| 14502 | s->it | -4.81 | 201 | 595 | *2* | *56* |
| 14503 | s->that | -5.09 | 201 | 315 | *1* | *34* |
| 14504 | dormouse-> the | -5.13 | 40 | 1642 | *1* | *35* |
| 14505 | queen->the | -5.17 | 75 | 1642 | *2* | *72* |
| 14506 | she->and | -5.78 | 552 | 872 | *1* | *55* |
| 14507 | was->she | -5.78 | 357 | 552 | *1* | *55* |
| 14508 | m->i | -5.86 | 63 | 545 | *1* | *58* |
| 14509 | was->it | -6.23 | 357 | 595 | *1* | *75* |

The μ-measure, in contrast, gives some collocations that are intuitively unlikely phrases consisting of high frequency words. In the case of "the-> the" there exists 1641 pairs that speak against that pairing, but it is hard to explain this in terms of local syntactic constraints. The negative scores seems to capture possible typographic errors.

**Table 2.2:** The last ten collocations by μ

| N | word pair | μ | $F_1$ | $F_2$ | $F_{12}$ |
|---|---|---|---|---|---|
| 14500 | she-> of | -3.37 | *552* | *513* | *1* |
| 14501 | to-> and | -3.54 | *729* | *872* | *2* |
| 14502 | a->i | -3.66 | *632* | *545* | *1* |
| 14503 | and-> of | -4.03 | *872* | *513* | *1* |
| 14504 | i->and | -4.12 | *545* | *872* | *1* |
| 14505 | she->and | -4.14 | *552* | *872* | *1* |
| 14506 | to->to | -4.28 | *729* | *729* | *1* |
| 14507 | and-> and | -4.80 | *872* | *872* | *1* |
| 14508 | i->the | -5.03 | *545* | *1642* | *1* |
| 14509 | the->the | -6.62 | *1642* | *1642* | *1* |

### Particle verbs

Particle verbs are hard to rank high for the μ-measure, because the individual frequencies of the particles are usually devastatingly high, and the frequency of the main verb in particle verb constructions are usually higher than average. The Δμ are, in general, good at finding such combinations if the order between the two words is fixed (**Table** 3.1).

**Table 3.1**: Some verb + particle (or negation)

| word pair | $N_\mu$ | $N_{\Delta\mu}$ | $\mu$ | $\Delta\mu$ |
|---|---|---|---|---|
| did->not | 3961 | 33 | 6.39 | 8.13 |
| seemed->to | 6818 | 54 | 4.80 | 7.64 |
| must->be | 4038 | 58 | 6.32 | 7.57 |
| looked->at | 5211 | 72 | 5.61 | 7.41 |

### Finding thematic phrases

*But what about finding Alice's friends?* Does the $\Delta\mu$ find the phrases that the text is *about* ($\approx$ thematic phrases)? To test this we chose some of the names of Alice's friends (**Table** 3.2).

We found that the *rank number* that $\Delta\mu$ delivers is higher than the rank number for the $\mu$-measure for all the checked friends. This is due to the frequency effects discussed above.

**Table 3.2**: Alice's Friends

| word pair | $N_\mu$ | $N_{\Delta\mu}$ | $\mu$ | $\Delta\mu$ |
|---|---|---|---|---|
| mock->turtle | 1517 | 12 | 8.86 | 9.13 |
| march->hare | 1003 | 28 | 9.65 | 8.28 |
| white->rabbit | 1637 | 47 | 8.62 | 7.78 |
| cheshire ->cat | 1360 | 473 | 9.04 | 5.64 |
| the->queen | 8519 | 831 | 4.00 | 5.17 |
| the->dormouse | 8841 | 832 | 3.86 | 5.13 |
| the->king | 8463 | 2954 | 4.03 | 3.63 |

### What is lost

There are obviously good phrases that $\mu$ rates higher than $\Delta\mu$. These usually consists of two words that are uncommon in the sample. Some idioms are of this kind. The $\Delta\mu$ needs to find more examples of collocations with the exact ordering between the consti–tuents to rate the collocation high ( **Table** 3.3).

**Table 3.3**: Some collocations with $N_\mu < N_{\Delta\mu}$

| word pair | $N_\mu$ | $N_{\Delta\mu}$ | $\mu$ | $\Delta\mu$ |
|---|---|---|---|---|
| yer->honour | 172 | 705 | 12.7 | 5.32 |
| young->lady | 230 | 1073 | 12.4 | 4.91 |
| guinea->pigs | 398 | 645 | 11.6 | 5.32 |
| rose->tree | 459 | 1114 | 11.3 | 4.91 |
| fast->asleep | 460 | 1115 | 11.3 | 4.91 |
| note->book | 462 | 2501 | 11.3 | 4.32 |
| raving->mad | 597 | 2500 | 10.8 | 4.32 |
| cheshire->cats | 1925 | 4468 | 8.23 | 3.32 |

### Adding memory

We have also done some experiments with adding memory to the method. A 'memory' could, for example, extend 10 words after each word. All words following within a distance equal to the size of the memory were collected. Adding a memory allowed the model to detect shared information of words that was further apart (for example "*pack* of *cards*" or "*boots* and *shoes*".

The memory introduced false collocations: e.g., "grammar-> mouse". The context was:

*"Alice thought this must be the right way of speaking to a mouse: she had never done such a thing before, but she remembered having seen in her brother's Latin* ***Grammar,*** *'A* ***mouse****--of a* ***mouse****--to a* ***mouse****--a* ***mouse****--O* ***mouse****!'"*

This context gave up to 5 collocations for "grammar" followed by "mouse", and therefore rated "grammar-> mouse" very high.

Otherwise, words that happened to be near a word without being statistically related to the word were usually rated low. The $\mu$ gave clearly better results on finding related phrases than the $\Delta\mu$, with the model with the 'memory'.

With the memory, the $\Delta\mu$ ordered the pairs closer to the original raw-frequency ordering the more 'memory' was present. The experiment with the memory was useful because it showed that this was not worth doing for $\Delta\mu$, but likely worth doing for $\mu$.

## CONCLUSIONS

### Possible usefulness

The higher sensitivity to local constraints in the temporal ordering could be used in a parser for finding local phrases. This might also have its implications for language acquisition. It could be tested if language learners make mistakes that could be explained by the statistical connectivity between words. Further research is needed on how the measure of connectivity behaves on phrase boundaries.

Areas where phrase finding could be useful include: text-to-speech (phrase intonation), machine translation (translation of compounds), and in information retrieval: phrase transformation of high frequency terms into medium frequency terms with a better discrimination value (Salton & McGill, 1983).

### Characteristics

The $\mu$–measure is good at estimating *global* correlations in a document or collection of documents (Wettler & Rapp, 1989). This could be used for capturing contextual and pragmatic constraints in a text. Other methods exist that are good, perhaps even better, at capturing for example synonymy.

Linear least square mapping (Yang & Chute 1992) is one method that has shown to be promising on capturing very good mappings between, in their case, symptoms and diagnosis. The same technique could be used for mapping a text to its abstract. The draw-back of these methods is their inherent parallel structure which makes it hard to account for the ordering that natural language requires.

The $\Delta\mu$- measure, on the other hand, is a local measure, that seems to capture dependencies in the temporal ordering of the language. It is hard to draw any definite conclusions from the analysis of only one text, but we have seen how the two proposed measures react to the frequencies of individual words, as well as the frequencies of word pairs. Taking into account the ability of $\Delta\mu$ to find dependencies in the temporal ordering, we think it is a more relevant measure than $\mu$ for several aspects of natural language processing, but not all.


## Acknowledgements

Thanks to the people at my department: especially Barbara Gawronska.


## REFERENCES


Belew, R. K., 1989, Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. In: Proc. SIGIR 1989, pp. 11-20, Cambridge, MA.

Fox, C., 1992, Lexical Analysis and Stoplists, In: Frakes, W. B., & Baeza-Yates, R., Information Retrieval, Prentice Hall, NJ.

Horne, M. & Johansson, C., 1991, Lexical Structure and accenting in English and Swedish restricted texts. *Working Papers (Dept. of Ling., U. of Lund, Sweden) 38: 97-114.*

Horne, M. & Johansson, C., 1993, Computational tracking of 'new' vs. 'given' information: implications for synthesis of intonation, In: Granström, B. & Nord, L. (Eds.) Nordic Prosody VI - papers from a symposium, Almquist & Wiksell International, Stockholm, Sweden.

Meisel, J., Clahsen, H., and Pienemann, M., 1981, On determining developmental stages in second language acquisition. *Studies in Second Language Acquisition 3, 2*, pp. 109-135.

Mitjushin, L. 1992, High Probability Syntactic Links, *Proceedings of the fifteenth International Conference on Computational Linguistics*, pp. 930-934.

Salton, G., & McGill M. J., 1983, *Introduction to Modern Information Retrieval*, McGraw-Hill Computer Science Series

Shannon, C. E., 1951, Prediction and Entropy of Printed English, Bell Systems Technical Journal, Vol. 30, No. 1, January 1951, pp. 50-65. (quoted in Salton & McGill)

Steier, A. M. & Belew, R. K., 1991, Exporting phrases: A statistical analysis of topical language. In Casey, R. & Croft, B., (Eds.), 2nd *Symposium on Document Analysis and Information Retrieval.*

Wettler, M. & Rapp, R., 1989, A connectionist System to Simulate Lexical Decisions in Information Retrieval, In: Pfeifer & al. (Eds.), *Connectionism in Perspective*, North-Holland

Yang, Y. & Chute C. G., 1992, A Linear Least Squares Fit Method for Information Retrieval from Natural Language Texts, *Proceedings of the fifteenth International Conference on Computational Linguistics*, pp. 447-453

Yarowsky, D., 1992, Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, *Proceedings of the fifteenth International Conference on Computational Linguistics*, pp. 454-460

Project Gutenberg, Illinois Benedictine College, send the message: *"send gutenberg catalog"* to *'almanac@oes.orst.edu'* for more information.

Carrol, L. *Alice's Adventures in Wonderland*, The Millennium Fulcrum Edition 2.9

# Information Retrieval & Extraction