# A Simple Categorisation Model of Anaphor Resolution

Andrew J. Stewart (Andrew.Stewart@Unilever.com)
Frederic Gosselin (gosselif@psy.gla.ac.uk)
Department of Psychology,52 Hillhead Street,
University of Glasgow, Glasgow G12 2QQ,
United Kingdom.

## Abstract

In this paper we examine the way in which approaching the task of anaphor resolution as a categorisation problem can shed light on the possible mechanisms underlying pronoun resolution. We formulate a model of anaphor resolution data within SLIP (Strategy Length & Internal Practicability) (Gosselin & Schyns, 1997, 1999), a general categorisation framework. We chiefly focus on pronominal anaphors in this paper but we also report the results of modelling repeat name anaphor reading time data collected by Stewart, Pickering and Sanford (in press). The success of adopting the redefinition of anaphor resolution as a categorisation problem suggests that problems faced by the cognitive system that have been considered unique to particular processing domains might be understood at a more cognitively general level.

## Introduction

In this article we bring together work on categorisation and work on psycholinguistics. We adopt a particular psycholinguistic phenomenon as a case study and examine it within a categorisation framework. We illustrate what a categorisation perspective can offer psycholinguistics in terms of theoretical apparatus. We examine the performance of a model formulated within the SLIP (Strategy Length & Internal Practicability) categorisation framework (Gosselin & Schyns, 1997, 1999), and show that it can account for human behaviour in pronoun resolution, a problem common in language processing.

We begin by reviewing existing work on pronoun resolution. Then we move on to our proposal which redefines the task of pronoun resolution as a categorisation problem. Following this we turn to outlining the SLIP framework. Finally, we discuss the consequences of redefining pronoun resolution as a categorisation problem and examine the correspondence between our model's predictions and experimental data.

## Existing Psycholinguistic Work on Pronoun Resolution

Anaphors are expressions that refer back to characters mentioned in a text. One example of an anaphor is a pronoun. Consider the fragment of sentence (A) up to but including the pronoun 'he'.

(A) John blamed Bill because he had damaged John's car.

This pronoun could refer to either character. Based on the information conveyed by the pronoun itself, the only restriction is that it refers to a singular male character. As both potential antecedents match on these features the sentence could plausibly continue like sentence (A) or (B):

(B) John blamed Bill because he didn't really like Bill.

In (A) the pronoun is coreferential with the character 'Bill', while in (B) it is coreferential with the character 'John'. There are a number of cues available in the text to facilitate the process of identifying the appropriate pronominal referent.

### Grammatical role cues

One cue is the grammatical positions occupied by the potential antecedents. The word 'John' occupies the grammatical subject position, while 'Bill' occupies the grammatical object position. A number of psychological theories, e.g. Subject Assignment Strategy (Stevenson, Nelson, & Stenning, 1995) and Parallel Function Strategy (Sheldon, 1974), predict a preference to interpret the referentially ambiguous pronoun in the above examples as coreferential with the grammatical subject (although for different reasons).

Note that in the examples discussed in this paper the character occupying the grammatical subject position is also the first mentioned character. Gernsbacher (Gernsbacher & Hargreaves, 1988; Gernsbacher, 1989) proposed that the first mentioned character occupies a privileged position in the reader's discourse model. A similar first mention privilege has been observed in other tasks (e.g. Neath, 1993; Neath & Knoedler, 1994). One of the consequences of the first mention preference found in language comprehension is that later in a sentence it is relatively easy to refer to the first mentioned character.

### Gender cues

Additional to grammatical position information, other cues may also be present. Consider sentences (C) and (D) below.

(C) John blamed Mary because she broke the window.
(D) John blamed Mary because he was in a bad mood.

The gender differentiation between the two characters serves as an additional (strong) cue as to which character the pronoun can refer. However, even under conditions where gender information can unambiguously identify the appropriate pronominal referent, there is much evidence to suggest that the system does not immediately take advantage of this (Stevenson & Vitkovitch, 1986; MacDonald & MacWinney, 1990; Tyler & Marslen-Wilson, 1982). It appears that gender information is treated simply as another cue, not in any way qualitatively distinct from other factors.

## Semantic cues

A particularly strong semantic cue known as implicit causality (Garvey & Caramazza, 1974) can also facilitate interpreting the pronoun. Implicit causality is a property associated with a particular set of verbs which influences processing of the pronoun in constructions such as 'John blamed Bill because he...'. It is manifested as a bias to interpret the pronoun as consistent with the implied locus of cause underlying the described event; such as the action of 'blaming' in this example. 'Blame' is classed as an NP2 biasing verb as it biases toward the character occupying the second Noun Phrase as the causal locus. Similarly there are also verbs such as 'fascinate' which bias toward the first Noun Phrase.

The explicit cause information contained in the subordinate clause (e.g. 'broke the window') is an important disambiguating cue. In Example (B) the fragment 'didn't really like Bill' indicates that the pronoun should be interpreted in a manner inconsistent with the implicit causality bias. The causality congruency effect (Garvey & Caramazza, 1974; McDonald & MacWhinney, 1995) is the finding that it takes longer to read a sentence where the implicit cause and explicit cause conflict than when they are consistent with each other.

So then, the cues available to aid identification of a pronoun's referent include order of mention, implicit cause, gender and explicit cause. Given the restriction that gender and explicit cause must agree, the set of all possible combinations of cues has a cardinality of 8. This total set is shown in Table 1 with example sentences exhibiting those features and with the mean reading times associated with reading the disambiguating fragment, i.e. the explicit cause (Stewart, Pickering & Sanford, in press).

Compared to the large body of work proposing and investigating possible parsing mechanisms, there are relatively few formal theories of pronoun resolution.

## Centering Theory

An adequate explanation of a process requires reference to a possible formal mechanism underlying that process and, for pronoun resolution, must take into consideration factors such as gender agreement and implicit causality verb biases. Centering Theory (Gordon, Grosz & Gilliom, 1993) is the best articulated theory in the literature. Centering proposes that utterances have associated with them a set of forward and a set of backward looking centres. The forward looking centre contains as its members entities, one of which forms the referential link between one utterance and the next. Factors such as the grammatical role of the characters in a text influence the ordering of the prominence of each of these entities. The backward looking centre of an utterance contains one member; the entity used to maintain reference between that utterance and the one preceding. Centering theory is a descriptive theory, rather than a processing theory, in as much as it describes the nature of the referential cohesion between units of a text. Although it describes what information might be used to facilitate pronominal reference resolution, it doesn't formalise how that information is used. This is hardly surprising as the theory originally grew out of work in Artificial Intelligence and so was never designed as a psychological model. How might a formal psychological model of pronoun resolution be arrived at? We propose that a possible way in which to arrive at a formal model of pronoun resolution is to make the explicit analogy between the problem faced by the processor in pronoun resolution and the problem faced by the processor in tasks of categorisation. In fact, at an important computational level we believe these problems are one and the same. There are many formal categorisation models and we believe that one in particular can be reinterpreted as a formal model of pronoun resolution.

# Mapping the problem of pronoun resolution onto that of categorisation

Let us return to Example (A), repeated below,

(A) John blamed Bill because he had damaged John's car.

The problem upon encountering the pronoun 'he' in this sentence can be understood as one of deciding of which category it is a member: should it be interpreted as a member of the set of expressions referring to the character 'John' or as a member of the set of expressions referring to the character 'Bill'? Furthermore, as we have discussed in above, this decision process is guided by explicit cause (and by gender, when it is relevant) and, to a lesser extent, by first mentioned character and by implicit causality information; these cues can be treated as features because they are discriminable parts of sentences that may be diagnostic with respect to the pronominal referent. Thus, a strong analogy can be made between problems of pronoun resolution and problems of categorisation. We shall study this parallel more thoroughly in the next section.

## A Categorisation Mechanism

SLIP (Strategy Length & Internal Practicability) was originally developed to model the results of experiments examining basic-levelness (Gosselin & Schyns, 1997, 1999). In this section we informally describe the SLIP framework and suggest how it can be used to model performance when faced with the type of categorisation problem required in identifying a pronominal referent. We provide a more complete treatment of this model in the Appendix.

We believe that pronoun resolution can be construed as a two-stage categorisation process. In the first stage, a hypothesis as to which referent is the most likely is generated. This is followed by the testing of this hypothesis. In the first stage, a SLIP categoriser extracts features randomly from the first half of the sentence. As soon as one critical features is selected, a hypothesis is formulated. We believe that the first stage is informed by

Table 1. Total set of feature combinations with example sentences, reaction times reported in Stewart, Pickering & Sanford (in press), Experiment 4 and theoretical predictions of our categorisation model.

| Sentence | Features | | | | | | | RT | Prediction |
|---|---|---|---|---|---|---|---|---|---|
| | F | NP1 | NP2 | G1 | G2 | CH1 | CH2 | | |
| | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1695 | 3.511 |
| (1) | John fascinated Mary because he was very interesting. | | | | | | | | |
| | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1980 | 9.851 |
| (2) | Mary fascinated John because he was easily interested. | | | | | | | | |
| | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1983 | 7.146 |
| (3) | John fascinated Bill because he was very interesting. | | | | | | | | |
| | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 2234 | 20.864 |
| (4) | John fascinated Bill because he was easily interested. | | | | | | | | |
| | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1769 | 6.681 |
| (5) | John blamed Mary because he was in a bad mood. | | | | | | | | |
| | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1641 | 6.681 |
| (6) | Mary blamed John because he broke the window. | | | | | | | | |
| | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1893 | 14.005 |
| (7) | John blamed Bill because he was in a bad mood. | | | | | | | | |
| | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1919 | 14.005 |

the first mentioned character and the implicit causality information. Order of mention is relatively salient and trivially recovered from the input. Au (1986) demonstrated that implicit causality information is also a very salient property. Both order of mention information and implicit causality contain some degree of uncertainty but they are also both useful predictors as to which way a sentence is going to continue (Garvey, Carmazza & Yates, 1975). The first mentioned character feature (F) can lead only to hypothesis_1, i.e. the hypothesis that the first referent is the pronominal referent. The implicit causality information, however, favours hypothesis_1 if the NP1 biasing implicit causality feature (NP1) is present in the sentence and hypothesis_2 (the hypothesis that the second mentioned character is the pronominal referent) otherwise.

Consider again the first portion of our example sentences (1) and (5) in Table 1:

(1) John fascinated Mary because he...
(5) John blamed Mary because he...

In the first case, the probability that hypothesis_1 will win is 1 because the two diagnostic features (first mention and implicit causality) both suggest that hypothesis_1 is appropriate. This is true of the first four example sentences in Table 1. For sentence (5) however, the probability that hypothesis_1 will win is only .5 as the two features contradict each other. This is true of example sentences (5)-(8).

The hypothesis that was adopted in the first stage and the diagnosticity of gender both influence which verification strategy will be adopted in the second stage.

Suppose, for instance, that a categoriser is presented example sentence (1) from Table 1:

(1) John fascinated Mary because he was very interesting.

At the end of stage one, the categoriser knows that gender information is relevant and it makes the hypothesis that 'John' is the correct referent (i.e. hypothesis_1). The extraction of either feature G1 or feature CH1 in the rest of the sentence verifies this hypothesis.

SLIP postulates a categoriser with a feature-extraction mechanism with a stochastic component. It is thus very likely that some features that are picked up by the categoriser are noninformative. For sentence (1), hypothesis_1 will ultimately be verified but this can take time. In the SLIP framework it is simple to compute the number of features, on average, that will be needed to be picked up for the categoriser to reach a decision (see Appendix). This is the measure reported in the simulation. The predictions of our model for all the sentences are shown in Table 1 together with reading time data reported in Stewart, Pickering and Sanford (in press).

Let us contrast the treatment of sentence (1) with one identical on all points except for gender diagnosticity. A categoriser is presented with sentence (3) from Table 1:

(3) John fascinated Bill because he was very interesting.

At the end of the first stage, hypothesis_1 is generated and gender information is known to be nondiagnostic. We thus have one nondiagnostic gender feature and one diagnostic CH1 feature in this case (i.e. CH1). In the terminology of the SLIP framework, this sentence has less redundancy than sentence 1. After a while, hypothesis_1 is also verified, but it takes longer to verify it in sentence (3) than in sentence (1) because of the lower redundancy of diagnostic information.

We now compare the first two situations with a third one in which the hypothesis formulated at the end of stage 1 is rejected in stage 2. A categoriser is shown example sentence (2) from Table 1:

(2) Mary fascinated John because he was easily interested.

At the end of stage 1, hypothesis_1 is proposed and gender is known to be diagnostic. This is similar to the outcome of stage 1 for sentence (1). Either G1 or CH1

would verify the hypothesis. Neither is present in the second portion of sentence (2) as the explicit cause information points to the second mentioned character (CH2). Thus, hypothesis_1 needs to be rejected and hypothesis_2 accepted. In the SLIP framework it is possible to compute a stop criterion based on an acceptable error rate so that if this criterion is reached, a revision of the hypothesis is made, i.e. the alternate hypothesis is adopted. In our simulation we have set the stop criteria at 11%, the error rate observed by Stewart, Pickering and Sanford (in press) (Experiment 4). Rejection of a hypothesis takes longer than verification of that hypothesis.

For sentences (5)-(8) from Table 1, the situation is slightly more complicated. Half the time hypothesis_1 is selected in stage 1; half the time, hypothesis_2 is selected. The average number of features that will be needed to be extracted before a decision can be made is the mean of that measure for the two possibilities. Take, for instance, example sentence (5) from Table 1:

(5) John blamed Mary because he was in a bad mood.

When hypothesis_1 is proposed, the treatment of sentence (5) becomes equivalent to example sentence (1) already discussed; when hypothesis_2 is elected, however, its treatment becomes equivalent to example sentence (2). So, the average number of features extracted before a decision is reached in sentence (5) is the mean of that in sentences (1) and (2). Arriving at a decision for sentence (5) is slower than (1) but faster than (2).

Stewart, Pickering and Sanford (in press) report the results of three further experiments examining the processing of anaphors in the context of sentences containing cues identical to the ones present in Experiment 4. The most important difference between those experiments and their Experiment 4 is that, while the anaphors in Experiment 4 are all pronouns, those in the remaining experiments are a mixture of ambiguous pronouns and unambiguous repeat names. In this paper we argue that the case of anaphor resolution can be reformulated as one of categorisation. Our main focus has been on the processing of anaphoric pronouns. To strengthen our argument, we need to show that our model also accounts for the processing of other types of anaphor. In addition to modelling Experiment 4 from Stewart, Pickering and Sanford (in press), we also modelled their Experiments 2 and 3 (deep processing condition). The raw Pearson correlations between the models' best predictions and the experimental data are .884 ($p < .05$; best predictions: 1.12286, 1.11834, 1.12060, 1.12060, 1.15261, 1.28229, 1.25107, 1.25107 in the order of Stewart, Pickering, & Sanford's Table 1), .817 ($p < .05$; best predictions: 1.20796, 1.64386, 1.42591, 1.42591, 1.38960, 2.02429, 1.69340, 1.69340 in the order of Stewart, Pickering, & Sanford's Table 1), and .816 ($p < .05$), respectively, for Experiments 2, 3, (deep-processing condition) , and 4. So, not only can our model correctly predict the reading time data associated with processing pronouns reported in Stewart, Pickering and Sanford (in press), it can also correctly predict the reading times associated with the processing of more general anaphoric expressions.

## Discussion

Our categorisation function explains the first mention effect (Gernsbacher & Hargreaves, 1988; Gernsbacher, 1989), the causality congruency effect (Caramazza, Grober, Garvey & Yates, 1977; Ehrlich, 1980; Garnham, Oakhill & Cruttenden, 1992), and the effect of gender diagnosticity (Caramazza et al, 1977; Garnham et al, 1992) reported in the psycholinguistic literature. As outlined above, the first mention privilege is the finding that the first mentioned character is easy to later refer to within the sentence in which it appears. By considering the first mentioned character as 'special', and by associating a feature with it, SLIP performs more quickly when this character is the pronominal referent than when it is the second mentioned character. In other words, our model predicts that pronoun resolution is relatively straightforward when a pronoun refers to the first mentioned character. Our model also accounts for the causality congruency effect. It predicts that pronouns are more difficult to resolve when they occur in a sentence containing an NP1 implicit cause and an NP2 explicit cause. Our model predicts that the causality congruency effect will not be found for NP2 implicit cause verb conditions where the explicit cause is NP1. This is because the first mention privilege allows some difficulty that arises as a result of the implicit causality inconsistency to be overcome. In other words, our model predicts that, all other things being equal, the causality effect is asymmetrical. Although the causality congruency effect has been widely reported in the literature (McDonald & MacWhinney, 1995), possible accounts of its asymmetrical nature have never been provided. Finally, our model predicts that it should be easier to identify a pronoun's antecedent when gender information differentiates between possible referents (Caramazza et al, 1977; Garnham et al, 1992). Additionally, it also offers a computational explanation for why this is the case. In light of the close correspondence between our model's predictions and well-established psycholinguistic phenomena it is clear that not only does our categorisation function successfully characterise human performance on tasks of anaphor resolution, it also provides an explanation at the level of categorisation with respect to why this pattern of performance arises.

The success of SLIP on tasks as (apparently) diverse as anaphor resolution and basic level categorisation suggests that other types of cognitive tasks may also benefit from their reinterpretation as categorisation problems. Understanding the degree to which computational problems faced by the cognitive system in specific processing domains can be interpreted as specific instances of more general problems allows for the proposal of mechanisms of greater explanatory power than those currently suggested in (for example) the literature on anaphor resolution.

## References


Au, T.K. (1986). A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language*, **25**, 104-122.

Caramazza, Grober, Garvey, and Yates (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behaviour*, **16**, 601-609.

Ehrlich, K. (1980). Comprehension of pronouns. *Quarterly Journal of Experimental Psychology*, **32**, 247-255.

Garnham, A., Oakhill, J. and Cruttenden, H. (1992). The role of implicit causality and gender cue in the interpetation of pronouns. *Language and Cognitive Processes*, **7**, 231-255.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, **5**, 459-464.

Garvey, C., Caramazza, A. and Yates, J. (1976). Factors underlying assignment of pronoun antecedents. *Cognition*, **3**, 227-243.

Gernsbacher, M.A. (1989). Mechanisms that improve referential accesss. *Cognition*, **32**, 99-156.

Gernsbacher, M.A., & Hargreaves, D.J. (1988). Accessing Sentence Participants: The Advantage of First Mention. *Journal of Memory and Language*, **27,** 699-717.

Gordon, P.C., Grosz, B.J., & Gilliom, L.A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, **17**, 311-347.

Gosselin, F., & Schyns, P. G. (1997). Debunking the basic level. In M. G. Shafto & P. Langley (Eds.), Proceedings of the nineteenth annual conference of the Cognitive Science Society (pp.277-282). New Jersey: Lawrence Erlbaum Associates, Publishers.

Gosselin, F. & Schyns., P.G. (1999, submitted). A new formal model of basic-level categorization and recognition, and its testing.

MacDonald, M.C., & MacWhinney, B. (1990). Measuring inhibition and facilitation from pronouns. *Journal of Memory and Language*, **29**, 469-492.

McDonald, J.L., & MacWhinney, B. (1995). The time course of pronoun resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language*, **34**, 543-566.

Neath, I. (1993). Distinctiveness and serial position effects in recognition. *Memory & Cognition*, **21**, 689-698.

Neath, I., & Knoedler, A.J. (1994). Distinctiveness and serial position effects in recognition and sentence processing. *Journal of Memory and Language*, **33**, 776-795.

Sheldon, A.L. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behaviour*, **13**, 272-281.

Stevenson, R.J., & Vitkovitch, M. (1986). The comprehension of anaphoric relations. *Language and Speech*, **29**, 335-357.

Stevenson, R.J., Nelson, A.W.R., & Stenning, K. (1995). The role of parallelism in strategies of pronoun comprehension. *Language and Speech*, **38**, 393-418.

Stewart, A.J., Pickering, M.J., & Sanford, A.J. (in press). The time-course of the influence of implicit causality information: Focus versus integration accounts. *Journal of Memory and Language*.

Tyler, L.K., & Marslen-Wilson, W. (1982). The resolution of discourse anaphors: Some on-line studies. *Text*, 2, 263-291.

## Appendix

The gist of SLIP is both simple and intuitively appealing: a classifier with an imperfect pick-up mechanism serially cycles through one or many strategies test by test in an attempt to verify one of them. A strategy gives the procedure required to check whether an object is a member of a given category. More specifically, a strategy is a series of sets of redundant features. For instance, take example sentence (1) in Table 1 :

(1) John fascinated Mary because he was very interesting.

At the end of stage 1, hypothesis_1 (i.e. the hypothesis according to which the first mentioned character is the pronominal referent) is made and gender is known to be diagnostic. This translates into the following strategy: S1 = [{G1, NP1}]. This is a length 1 strategy because it has only one set of redundant features. All the strategies required for pronoun resolution are of length 1 although for SLIP this does not have to be the case (see Gosselin & Schyns, 1997, 1999). For the sake of simplicity our formal discussion is confined to length 1 strategies here. The set of redundant features in S1 contains all the features which can decisively verify hypothesis_1 in example sentence 1. Three other strategies are also used for the set of example sentences in Table 1: S2 = [{NP1}], S3 = [{G2, NP2}], and S4 = [{NP2}]. S2 is used when hypothesis_1 is made and gender is nondiagnostic; S3 is employed when hypothesis_2 is made and gender is diagnostic; and S4 is used when hypothesis_2 is made and gender is nondiagnostic.

In the SLIP framework, a strategy as a whole is verified whenever all sets of redundant features have been individually verified in a specific order. A set of redundant features has been verified as soon as a one of its features has been verified. For example, S1 is verified as soon as either G1or NP1 is verified. Given that a SLIP categoriser has a stochastic feature-pick-up mechanism, this verification habitually happens after a succession of misses. The probability of having t-1 successive misses is given by $(P\text{-}PQ)^{(t\text{-}1)}$ where $P$ is the probability of a random slip and $Q$ is the probability of a diagnostic slip, i.e. the cardinality of the set of redundant features divided by the total number of features in the shown sentence. We assume in this article that 10 features are present in sentences for the verification stage: gender information (sometimes diagnostic and sometimes not), explicit cause (always diagnostic), and eight nondiagnostic features such as verb tense (this number was arbitrarily chosen, but a different one would make little difference). The probability of a hit is simply 1 minus the probability of a miss. Thus, the probability that a certain strategy will be verified after t tests is:

$(P\text{-}PQ)^{(t\text{-}1)}[1\text{-}(P\text{-}PQ)]$.

This expression gives the Special Response Time Density Function (SRTDF) of a SLIP categoriser. It describes a geometric density function. The best fit between the data and our predictions is obtained with $P = 1$, meaning that features are gathered randomly.

The global measure reported in our simulations is the average number of features that have to be picked up before the categoriser reaches a decision (i.e. to verify or reject a strategy). We begin with the rejection case. If a categoriser has failed to verify a strategy after *t_stop* (*t_stop* = 1) feature pick-ups either the strategy does not apply, or the categoriser's extraction mechanism has until then slipped onto nondiagnostic features. As *t_stop* increases the second possibility becomes less and less likely. A classifier could thus conclude quite confidently that a strategy does not apply if it has reached *t_stop* pick-ups if beyond this point the probability that the strategy applies to the pronoun is smaller than some small constant probability *D*. Given *P*, *Q* and *D*, *t_stop* can be calculated easily:

$t\_stop = \log D / \log(P\text{-}PQ)$.

This equation is known as the inverse survival function of probability *D*. A categoriser using this method errs with a probability of *D* on negative trials (i.e. it rejects the hypothesis when it is correct with a probability of *D*). For the simulations *D* was set at .111, the subjects' mean error rate in Stewart, Pickering and Sanford (in press, Experiment 4). Note: this is not a free parameter. Consider example sentence (2). $Q = 2/10$. It thus takes our categoriser an average of 9.851 pick-ups before rejecting hypothesis_1 and thus accepting the alternative hypothesis_2.

Now that we know how to compute *t_stop*, we can calculate t_mean, the mean number of pick-ups required to verify positive trials (i.e. when a strategy is correct):

$$t\_mean = \frac{\sum_{t=1}^{t\_stop'} t.SRTDF}{\sum_{t=1}^{t\_stop'} SRTDF}$$

where *t_stop'* is simply *t_stop* rounded up to the next integer. Consider example sentence (1). Q = 2/10. We can thus use the *t_stop* calculated for example sentence (2); once rounded up it becomes 10. So, *t_mean* is equal to 3.511; it takes an average of 3.511 pick-ups for hypothesis_1 to be accepted in this case.