Chapter 1

Evaluating reference resolution

A guide to numeric measures

Andrei Popescu-Belis
University of California, San Diego

Key words: Reference, coreference resolution, quantitative quality measures

Abstract: Computer programs are increasingly capable of grouping together expressions of a discourse that denote the same entity. We regard the evaluation of this capacity as the comparison between a system's response and the one expected by the evaluators. We outline a theoretical framework for reference (§1) and another one for evaluation (§2), then analyze three existing quality measures (§3.2–3.4), one of which was used in the MUC evaluation campaign. We propose mainly two new measures, one based on the notion of core equivalence class (§3.5), and the other based on information theory (§3.6), both showing better theoretical coherence than the previous ones. We also examine two alternatives, the exclusive core classes (§3.5.4) and the distributional measure (§3.7). In addition, we study a series of generalizations to the main problem (§4), and provide the results of all measures on several texts (5).

Language-related tasks often require a certain degree of language understanding. Following a commonsense conception, understanding a linguistic message will be equated here with: (1) understanding what entities the message talks about; (2) understanding what the message says about these entities (their properties, relations, etc.) Our main goal is to estimate the capacity of a computer program to "understand references", that is, to keep track of the various entities that a linguistic message is about. We will first describe a broad framework for this phenomenon, and situate the problem of coreference within it.

1. A FRAMEWORK FOR REFERENCE USE

1.1 Referring acts

Let us suppose that the world or *environment* is made of *entities*, as for instance individuals, objects, abstract ideas, etc., which possess various *properties* and *relations*. The *agents* that use language have representational capacities allowing them to manage *mental representations of the entities* (henceforth, MRs). The notion of MR is versatile enough to accommodate various *referring cases*, i.e. situations in which the represented entity may be more or less specific, determined, unknown, generic, etc.

The speaker of *sender* produces a linguistic *message* and addresses it more or less directly to a hearer or *receiver*. The notion of *referring act* supposes the following:

1. For each utterance in the message, the sender has in mind or *activates* one or several MRs, together with one or more properties that concern the MRs.
2. For each MR that the speaker activates in their mind, a fragment of the utterance is uniquely related to the activation of that MR.
3. Upon reception of the utterance, each particular fragment activates one of the receiver's MRs, which may be old or just created.

The utterance fragment constitutes a *referring expression* (henceforth, RE). Condition (3) states that for an utterance to give rise to a referring act, it is necessary that the REs be understood as such and that the receiver activates one or more MRs upon reception. Activating an MR for a phrase that did not intend to activate one, or failing to activate an MR where an MR should have been activated is not misunderstanding reference, it is no referring at all. *Reference* is the link between an entity, an MR and an RE, be it in the sender or in the receiver.

1.2 Felicitous referring acts

Intuitively, we would be tempted to say that a referring act was felicitous, or that a *reference was understood*, if the receiver activates the "same" MR as the sender. Unfortunately, this straightforward criterion is neither tenable on theoretical grounds, nor applicable in practice.

First, it is probably not true that the sender and the receiver have comparable MRs, since their view of the world is different, as well as their "minds". Even when two MRs (sender *vs.* receiver) represent the same well-known entity, the properties that the MRs gather are probably not exactly the same. Also, suppose that the receiver divides the properties of a single entity between two MRs, being unaware of the identity of the two – e.g., someone not knowing that Marcus Aurelius was both a stoic philosopher and a Roman emperor. This is especially the case with computer programs. Which of the two MRs counts as "the correct one"? Finally, it is not always possible to access explicitly the complete structure of an MR, especially with human agents. The solution is then to ask questions about the MR that was activated by a particular RE, or simply activate again, as a sender, and check whether the *same* MR is again activated in the receiver.

> Therefore, finding out whether a referring act has been felicitous (i.e., *evaluating reference understanding*) is possible *only* through subsequent referring acts activating the same speaker MR. Evaluation is essentially performed on sets of REs, not on individual referring acts.

A way to estimate whether a referring act has been felicitous is to ask the hearer subsequent questions about the entity whose MR was supposed to be activated. Suppose for instance that A talks to B about a certain ski run, and wants to make sure that B thinks about the same one. After the first referring act (A: "I went down that long bumpy run", B: "Yeah, very bumpy"), A tries further references to the same entity (A: "You know, that steep looping run", B: "Yes, the one with three bends" *or* B: "How come? The bumpy one goes straight down"). Still a second "Yes" may not be enough to A, so the test may continue for a while, depending on how important the confusion may be (how many long bumpy steep looping runs there are).

Our approach is of course mainly concerned with the case when the receiver is a computer program (e.g., a text processing device), which cannot generally answer such questions. It is possible, but unpractical, to make explicit the receiver's MRs and then compare them with the

sender's MRs: this is uneasy especially if the understanding is not almost perfect. Another idea is to give the program the set of possible MRs in advance (a "reference" set) and ask it to pick the right MR for each referring act, instead of managing its own set. This is unpractical because we would like to see the program build its own MRs, especially for newly encountered entities.

Therefore, a tractable way to measure reference understanding is to use the series of receiver MRs activated by referring acts, and compare it to the series of sender MRs, not one by one but in terms of *correlation*. Two equivalent points of view will enable us to estimate reference understanding under the previous limitations. Understanding is correct if multiple referring acts activating a certain MR in the sender always activate *the same MR* in the receiver. Alternatively, in terms of referring expressions, understanding is correct if all the REs corresponding to a sender MR also correspond to a unique receiver MR.

1.3 Infelicitous referring acts

We introduce now two types of understanding errors. Suppose that after a first referring act, the sender produces a second one, which may activate either the same MR, or another one. The hearer also activates an MR, which may be the one activated for the first act, or another one. There are four possibilities, the two incorrect ones being shown *Figure 1*.

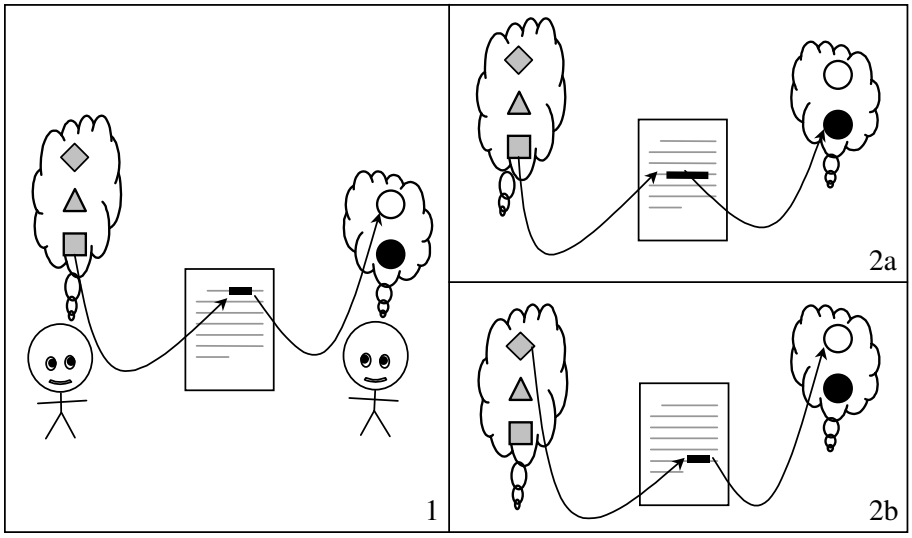

Figure 1. Two types of reference understanding errors: (1) first referring act, (2a,b) two further referring acts – (2a) r-error, (2b) p-error with respect to (1)

In (2a), the sender activated the same MR, while the receiver activated another one instead of activating the same one ("the receiver believed that the sender referred to another object"). This error introduces a rupture in the structure of MRs and we will call it an *r-error* ('r' is chosen for reasons that will appear later). It may also be viewed as a "missing link" between two referring acts or two REs. The fact that the activated MR already existed or was newly created is not relevant for r-errors.

In (2b), the sender activated a different MR on the second referring act, while the receiver re-activated the previous one, instead of switching to another one ("believed that the sender referred to the same object"). This second type, obviously quite different from the first one, puts together two referring expressions that shouldn't be associated, and we will call it a *p-error*

(again ‘p’ is chosen with intent). It may also be viewed as a “wrong link” between two referring acts.

A referring act may generate both types of mistakes simultaneously, provided that at least two referring acts have preceded it. The sender may have first activated MR_1 and the receiver MR_a, then the sender activated MR_2 while the received activated MR_b. A subsequent activation of MR_1 by the sender and of MR_b by the receiver would count for two errors, an r-error for not activating MR_a, and a p-error for wrongly activating MR_b. Activating a newly created MR_c would avoid the p-error.

1.4 Examples

Quantitative evaluation goes far beyond cases with two or three referring acts. The simple counting method described above encounters severe problems in the case where numerous MRs and REs are present. Suppose that the sender activates MR_1 *ten times consecutively*, and the receiver activates MR_a for the first referring act, and MR_b for the second, which counts as an r-error. What happens then if on the third referring act the receiver activates again MR_b? Is this another r-error, with respect to the first act? Or is this correct, with respect to the second act? What happens if, on the fourth referring act, the receiver activates MR_a again? Obviously, such a serial count is uneasy.

To get a better view of the counting options, suppose globally that out of the ten referring acts in which the speaker always activated MR_1, five acts activated in the speaker MR_a and five MR_b, regardless of the order. How many r-errors should we count? Possible answers are *five* (say all activations of MR_b) or *one* (as there may have been only one rupture that unduly created MR_b). The number of referring acts is also relevant: in this example, the receiver’s reaction would be considered quite good if the sender also activated lots of other MRs, and quite average if the sender activated only MR_1.

Finally, consider the following text (translated from a French guidebook), which we will use throughout the article as an example.

> The western peak$_{(1)}$ is 10,254 feet high. To reach it$_{(2)}$, follow for about 300 ft. a narrow passage$_{(3)}$ that$_{(4)}$ is often icy and slippery. This passage $_{(5)}$ starts right behind the southern peak$_{(6)}$ (9,742 ft.), which$_{(7)}$ is, as for it$_{(8)}$, much easier to reach. This second peak$_{(9)}$ is well visible, because it$_{(10)}$ is very prominent. In order to reach this small turret$_{(11)}$, constantly aim at it$_{(12)}$ from the large lower passage$_{(13)}$ (quite easy to climb). This one$_{(14)}$ is initially wide, but it$_{(15)}$ becomes narrower as it$_{(16)}$ runs up. Beware, this inviting cradle$_{(17)}$ is somewhat slippery too.

There are 17 referring acts, hence 17 REs. Quite obviously for the reader, the sender (author) successively activated four MRs: <western peak>, <upper passage>, <southern peak> and <lower passage>. The set of correct sets of REs corresponding to each MR is termed the *key* (*Table 1*, left column). A program that “understands” the text should build some sort of representation of the two peaks and the two passages. Alternatively, as far as reference understanding is concerned, the program should at least group the REs into sets that correspond to the same MR (cf. *Table 1*, right column, for a sample *response*).

Table 1. Key and response RE sets for the example test

Key: P_K	**Response: P_R**
K1: 1, 2	*R1*: 1, 2, 6, 7, 8, 9, 10
K2: 3, 4, 5	*R2*: 3, 4, 5, 11, 12, 13, 14, 15, 16
K3: 6, 7, 8, 9, 10, 11, 12	*R3*: 17
K4: 13, 14, 15, 16, 17	

Evaluating the response means answering the question: how far is the response from the key? That is, how far is the system's output on that given text from the expected output? Ideally, the answer should be a number (a score), so that responses on different texts and/or from different programs may be compared. As the last example shows, the score should be computed via an algorithm using the key and the response, as human judgment of this data is uneasy.

1.5 About coreference

We have defined successful reference transmission between a sender and a receiver as the constant co-activation of MRs. The last example however introduced the sets of REs corresponding to the same MR, in the sender or in the receiver. Quite logically, such REs are termed *coreferent*, and the sets of coreferent REs partition the set of all the REs into several *classes*. Then, successful reference transmission means also that the sender and the receiver have the same classes, under the hypothesis that they agree on the total set of REs.

Now, two REs corresponding to the same MR being coreferent, a lot of importance has been attached to the *coreference link* between them. The significance of such a link is problematic, since it is more likely that a human receiver will interpret REs depending on his/her MRs, and not directly on previous REs (except maybe for some pronouns). The use of coreference links has been one of the main problems in the early studies of evaluation, since there are many combinations of links that correspond to the same understanding (same classes, same co-activations). Given the links, the solution was to build the classes (transitively following all the links) and only afterwards proceed to evaluation.

1.6 Towards quantitative evaluation

Let us summarize the conclusions of this section. Understanding referring acts requires two capabilities: (1) detect referring acts, or referring expressions; (2) correctly activate the corresponding MRs. One should evaluate these capabilities separately, with the second one being the most relevant and the most difficult to evaluate (more about the first one in §4.3). We have described two error types, r-errors and p-errors, but their exact definitions vary according to each proposal (cf. §3). In addition to our previous description, ruptures (r-errors) have also been conceived as "missing coreference links", hence the name of *recall* errors borrowed from information retrieval (Van Rijsbergen 1979). Likewise, p-errors have been conceived as "wrong coreference links", hence *precision* errors[1]. While these terms may help understanding the problem, only the formulae given below constitute reliable definitions of what is exactly evaluated. These definitions are themselves subject to coherence criteria that we shall now describe.

[1] This conception is based however on coreference links, which have at least two problems: (1) different link configurations may in fact correspond to the same classes; (2) MRs activated only once have no links.

2. EVALUATION MEASURES FOR NLP

The second domain relevant to our present study – besides reference resolution – is the domain of NLP systems evaluation (Popescu-Belis 1999a). Within the black box / glass box opposition, our approach here favors the former, as we aim to evaluate a system's quality using only its output or response.

An evaluation *measure* is an algorithm that uses the input and the output data to produce a *single numeric score* representing the quality of the output (response) with respect to the desired processing of the input (key). When averaging the output quality over several inputs, an indication of the system's quality itself is obtained. An evaluation measure allows the comparison between various systems or between various states of a given system. In general, several quality indicators are measured on the output data, then integrated in a single score. An evaluation measure thus defines a mapping between the quality levels of an output (e.g. "perfect", "good", "average", "bad", "worthless") and a set of marks or ratings, either a discrete set, or here the [0%; 100%] interval. In theory, the "objective" mapping is the one determined by a large panel of human experts allowed to use in some detail the evaluated system. Our goal here is to propose formal measures that can be automatically computed using the system's output.

2.1 The MUC campaigns. The modularity of evaluation

Competitive evaluations have often been held as collective evaluation campaigns, e.g., MUC, TREC, TDT[2] (Hirschman 1998). Their main goal has been to compare the efficiency of different techniques on a given problem and to provide a snapshot of the current capabilities. The MUC campaigns were aimed at evaluating the capacity to "understand" short articles on a given theme, that is, to instantiate pre-determined templates with elements extracted from the articles[3] (Grishman and Sundheim 1996, MUC-6 1995, MUC-7 1998). Several subtasks were identified, as for instance labeling entity names, detecting coreference, instatiating the actor and attribute fields in the template. In order to evaluate each subtask, the corresponding correct answers (keys) had been created.

Coreference resolution, as a subtask, depends mainly on the identification of the REs and their correct tagging. In order to evaluate (co)reference resolution *exclusively*, the systems should start from the same correct RE set, as RE identification is strictly speaking another task. Evaluating coreference with the output of a system that has started from raw text means that the system could be unjustly penalized for its poor POS tagger or RE identification module, despite a strong reference resolution mechanism. However, this was not the MUC approach, and correct input data at every level was not created, modules being tested on results from previous modules.

[2] MUC: Message Understanding Conferences; TDT: Topic Detection and Tracking Project; TREC: Text Retrieval Conferences.

[3] Seven MUC evaluation campaigns and conferences have been organized (*http://www.muc.saic.com*). There were 18 participants at MUC-7 in 1998.

2.2 Coherence criteria for evaluation measures

The definition of an evaluation measure should try to capture in its formulae the judgment of human experts on the desired output of a program. No formal arguments can prove the exactitude of a measure, but there are some common sense criteria that a measure should be proved to satisfy[4]. Let us consider an evaluation measure that computes a score in the [0%; 100%] interval, using a *response* and comparing it to the *key* (or a set of keys).

1. **Upper limit criterion**: perfect responses (keys), and only them, should receive the maximum score of 100%.
2. **Lower limit criterion**: the worst responses, and only them, should receive the minimal score of 0%. The definition of the worst responses ("no processing" of the input) depends on the evaluators' considerations, and quite often cannot be described precisely. This criterion can be unfolded in two parts, (3) and (4).
3. **Direct lower limit**: all responses scoring 0% should be among the worst.
4. **Reciprocal lower limit**: all the worst responses should score 0%. In progressive terms, *the bad responses must receive low scores*. This criterion entails the following one.
5. **Low scores criterion**: the evaluation measure should be able to yield 0% scores, or at least low scores. Quite obviously, scoring 55% with a measure that never goes below 50% is not a significant performance.
6. **Relative indulgence / severity**: this is a comparison criterion between two measures, stating that m_1 is more indulgent (or lenient) than m_2 if it provides higher scores on a certain response domain. It is of course uneasy to prove such a property, except sometimes on particular domains. The notion of indulgence/severity should help choosing the most sensitive measure on the expected response domain, e.g., if responses are poor, chose a lenient one, and a severe one if they are good.

2.3 Combining elementary measures: *f-measure*

Quite often, the final score is a combination of several elementary measures, e.g., in our case, the number of r-errors and the number of p-errors. It is of course significant to keep both scores for a more precise evaluation (and display them using (x, y) coordinates), but sometimes a unique score is needed to summarize a system's performance. We may consider the average of the two scores, but a more common convention is to use the harmonic mean, or *f-measure*, defined as follows:

$$f-measure(r, p) = \frac{2}{1/r + 1/p}, \text{ or 0 if } r = 0 \text{ or } p = 0.$$

The harmonic mean has the advantage of being closer to the lower value of the two scores, all the more than this value tends to zero. In other words, if $r > 0$ is fixed and p tends to zero, then the *f-measure* tends to zero too – unlike the arithmetic mean – thus penalizing huge differences between r and p, or values close to zero. The f-measure can reach 0% if either r or p can do so, not necessarily at the same time.

[4] Fore a more thorough analysis of these problems, see (Popescu-Belis 1999a).

3. EVALUATION MEASURES FOR REFERENCE RESOLUTION

The definitions of recall and precision given at MUC-5 for (co)reference resolution have been significantly improved in a paper by M. Vilain *et al.* (1995), and subsequently used at MUC-6 and 7 (cf. 3.2). The main idea of the authors was to compute a score of missing and wrong coreference links that did not depend on the exact link configuration, but only on the RE sets. More recent studies, including ours, have shown that this measure is in some cases excessively lenient. Attempts to define a more accurate measure include those by R. Passonneau (1997) using the κ (*kappa*) factor (cf. 3.3) and by A. Bagga and B. Baldwin (1998a, 1998b) with the B^3 measure (cf. 3.4). We propose a method to compare RE sets based on the notion of core sets (cf. 3.5, and 3.5.4 for an extension), as well as a distributional method (cf. 3.7). We also apply information theory considerations in order to define referring information and a measure for its transmission (cf. 3.6).

In the following sections, we will provide synthetic formulae for all these measures, and analyze their compliance to the coherence criteria, and their relative indulgence. The score notation is based on three letters: the first one indicates the measure (*M*, *B*, *C*, *X* or *H*), the second one is recall or precision (*R* or *P*) and the last one is success or error (*S* or *E*, with $\_S + \_E = 100\%$). The κ and distributional measures are apart. Before describing the measures, we define some useful concepts and notations.

3.1 Theoretical prerequisites

The starting point for reference understanding evaluation is the set of all REs, the same for the sender and the receiver (but see §4.3 for a different view). Then we consider the distributions of REs into sets of REs that activated the same MR, first in the sender (key) then in the receiver (response). These sets contain in fact all the relevant information, whatever the theoretical grounds of the measure may be (constant co-activation of MRs, same RE sets, same coreference links). Sometimes the key and the response are defined by coreference links, so in this case the sets should be built using the transitive closure of the link sets.

The sets of REs activating the same MR (coreferent REs) are thus *equivalence classes* for the coreference relation, either for the sender or for the receiver. Indeed, an RE belongs to one and only one class (possibly a singleton class) and the classes form a *partition* of the RE set, key partition vs. response partition. Measuring the proximity between the key and response partitions of the same RE set is not a common mathematical problem. Set theory defines only the notion of being more or less fine-grained, or not comparable, but we will see that information theory provides some indirect results on partition comparison.

Let E be the set of REs, and let P_K be the key partition, that is, a set of subsets of E, $P_K = \{K_1, K_2, \ldots, K_n\}$, that are non empty, do not overlap, and recover E (equivalence classes) – cf. example in *Figure 2*. Likewise, we have the response partition $P_R = \{R_1, R_2, \ldots, R_m\}$. The sets or classes K_i and P_j may be singleton sets (cf. in Section 4.2 the case when singletons do not count for evaluation). The perfect answer corresponds to $P_R = P_K$, that is, for each K_i there exists R_j such as $R_j = K_i$. When this is not the case, it is useful to consider all the response classes that contain fragments of a given key class K. The *projection* of K on P_R is first defined as the set of fragments into which K is divided in the response partition:

(DEF.1) $\quad \pi(K) = \{A \mid \exists\, R_j \in P_R \text{ such as } A = K \cap R_j \text{ and } A \neq \varnothing\}$

The set of response classes that contain these fragments is:

(DEF.2) $\pi^*(K) = \{R_j \mid R_j \in P_R \text{ and } R_j \cap K \neq \varnothing\}$

Conversely, we define the projection of a response class R on P_K and the set of key classes containing the fragments:

(DEF.3) $\sigma(R) = \{B \mid \exists\, K_i \in P_K \text{ such as } B = R \cap K_i \text{ and } B \neq \varnothing\}$

(DEF.4) $\sigma^*(R) = \{K_i \mid K_i \in P_K \text{ and } K_i \cap R \neq \varnothing\}$

It follows that $\pi(K) \subset \mathit{Subsets}(K)$, $\pi^*(K) \subset P_R$, $\sigma(R) \subset \mathit{Subsets}(R)$ and $\sigma^*(R) \subset P_K$. Let us define the key coreference rate as $|E| / |P_K|$, i.e. the average number of REs per key equivalence class (cf. *Table 5* for examples). So, $|P_K|$ is the number of key classes or MRs activated by the sender and $|P_R|$ is the number of response classes or MRs activated by the receiver. Each key class K has at least one projection (itself) and at most $|K|$ (if it is completely fragmented), so the following inequalities hold:

(PROP.1) $1 \leq |\pi(K)| \leq |K|$ and $1 \leq |\sigma(R)| \leq |R|$, for all $K \in P_K$ and $R \in P_R$.

Let us illustrate these definitions on the sample test given above (Section 1.4), which has a key coreference rate of 17/4 = 4.25 ER/class (response: 5.67). There are four key classes and three response classes, both shown in *Figure 2*. K_1 and K_2 project onto P_R as single fragments, while K_3 and K_4 are both divided in two: $\pi(K_3) = \{\{6, 7, 8, 9, 10\}, \{11, 12\}\}$ and $\pi(K_4) = \{\{13, 14, 15, 16\}, \{17\}\}$. We have thus $\pi^*(K_1) = \{R_1\}$, $\pi^*(K_2) = \{R_2\}$, $\pi^*(K_3) = \{R_1, R_2\}$ and $\pi^*(K_4) = \{R_2, R_3\}$. In the same way, R_1 projects in two fragments, R_2 in three (shaded areas in *Figure 2*) and R_3 in only one. So, $\sigma^*(R_1) = \{K_1, K_3\}$, $\sigma^*(R_2) = \{K_2, K_3, K_4\}$, and $\sigma^*(R_3) = \{K_4\}$.

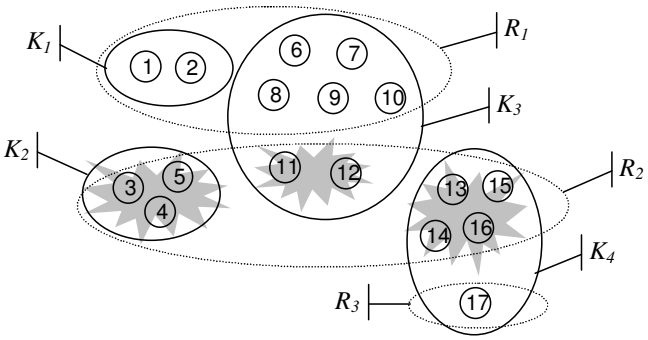

Figure 2. Key classes (solid line) and response classes (dashed line) for the example text in Section 1.4. Each RE is represented only once, as a circled number. Shaded areas represent $\sigma^*(R_2)$, the projection of R_2 on P_K.

3.2 MUC measure (M. Vilain *et al.*)

This measure defines the recall error for each key equivalence class K as the minimum number of links that are needed to reconnect all the projections of K on the response partition P_R (all the elements of $\pi(K)$). To compute the total recall error, the figures for each K are added and the sum is divided by the maximum possible value; then, recall success is 100% minus the error.

For instance, on *Figure 2*, the key classes K_1 and K_2 do not give rise to recall errors because they are not fragmented, $|\pi(K_1)|= |\pi(K_2)|=1$. However, K_3 does, as it is divided among two response classes. The MUC measure estimates that a single coreference link has been missed (among six, as $|K_3|=7$), say between ER_{10} and ER_{11}. Also, for K_4, a single link among four has been supposedly missed, say between ER_{16} and ER_{17}. The MUC recall error is thus $MRE = (0+0+1+1) / (1+2+6+4) \approx 15\%$, hence $MRS \approx 85\%$.

We have derived an explicit formula for this scoring algorithm described by M. Vilain *et al.* (1995), and we use the formula as a definition for the MUC score:

(DEF.5) $$MRS(P_R, P_K) = \frac{|E| - \sum_{K \in P_K} |\pi(K)|}{|E| - |P_K|}$$ and $MRS = 1$ if $|E| = |P_K|$

Conversely, the number of "wrong links" that figure in a response class is computed using its projections on the key partition P_K, hence the symmetrical formula for precision:

(DEF.6) $$MPS(P_R, P_K) = \frac{|E| - \sum_{R \in P_R} |\sigma(R)|}{|E| - |P_R|}$$ and $MPS = 1$ if $|E| = |P_R|$

Note that the minimum number of links necessary to form all the key classes (P_K) is $|E| - |P_K|$, and $|E| - |P_R|$ for all response classes (P_R). When either is zero, we have chosen here coherent conventions, as the cases were not described by the authors. For instance, $|P_K| = |E|$ means that there is no coreference in the key (all classes are singletons, each RE activates a different MR), so there can be no recall error. Conversely, $|P_R| = |E|$ means that the response is made of singleton MRs (no resolution, in fact), so there can be no precision error. It can be shown however (using the next result) that in both cases the *f-measure* equals zero, unless $|P_K| = |P_R| = |E|$ when *f-measure* is 1.

The following result is visible in the MUC scoring reports, and we prove it in the Appendix – it is also a common result in information retrieval:

(PROP.2) the upper parts of the *MRS* and *MPS* fractions are equal

What about coherence criteria? The following results prove that the upper limit criterion (1) is satisfied ("∃!" means "there exists a unique…"):

(PROP.3) $MRS = 100\% \Leftrightarrow \forall K \in P_K,\ \exists!\ R \in P_R$ such as $K \subset R$
$MPS = 100\% \Leftrightarrow \forall R \in P_R,\ \exists!\ K \in P_K$ such as $R \subset K$
f-measure $= 100\% \Leftrightarrow MRS = MPS = 100\% \Leftrightarrow P_K = P_R$

The lower limit criteria (2 to 5) are more difficult to study, as it is not easy to determinate which responses are noted 0%. The MUC measure can reach 0% scores, so the low scores criterion (5) is satisfied; for instance, no resolution at all (all response classes are singletons) leads to $MRS = 0\%$ and $MPS = 100\%$, so *f-measure* $= 0$. However, the reciprocal lower limit criterion (4) – bad responses receive low scores – seems violated, as we can find poor responses that receive high scores. For instance, if the system groups all REs into a single class:

(PROP.4) $P_R = \{E\} \Rightarrow MRS = 100\%$ and $MPS = \dfrac{|E| - |P_K|}{|E| - 1}$

Therefore, a very simple strategy, i.e. a poor response, obtains a non-zero score, actually a score that increases when the key coreference rate is important. We have also proved the following inequalities (cf. Appendix):

(PROP.5) $MRS \geq \dfrac{|E| - |P_K| \cdot |P_R|}{|E| - |P_K|}$ and $MPS \geq \dfrac{|E| - |P_K| \cdot |P_R|}{|E| - |P_R|}$

The graphic representation of the lower limits (*Figure 3*) shows that if $|E| / |P_K| >> 2$ (high coreference rate) then a response with $|P_R| < |E| / |P_K|$ (few classes) obtains a positive score. So, for texts with high coreference rates, the MUC measure does not satisfy the lower limit criteria[5]

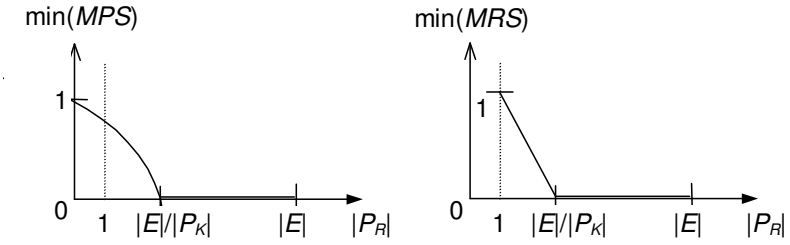

Figure 3. Lower limit of the MUC measure depending on the number of response classes (left, precision and right, recall)

3.3 Computing the κ factor (R. Passonneau)

The object of Passonneau's study is the measure of inter-annotator agreement. Instead of the key and the response, two "key" partitions have to be compared. Even if the agreement is good in general, it is not perfect, and the MUC measure seems too lenient to measure the small disagreement. Considerations inspired from the *kappa* measure (Krippendorff 1980) are applied in order to estimate the probability of random agreement, and find out how much above random is the actual agreement. This of course applies also to the comparison between a key and a response.

Table 2. Probabilities of agreement on the present/absent links between two partitions

	link exists in P_{R1}	link does not exist in P_{R1}
link exists in P_{R2}	*a*	*b*
link does not exist in P_{R2}	*d*	*c*

The idea is to represent four quantities in the 2x2 table shown above, viz., the probabilities that a link be present or not at the same time for the two annotators. Passonneau notes that it is not the exact links that matter, so it is not possible to count *a*, *b*, *c* and *d* directly, but they may be computed using the following formulae:

[5] Another lower limit for *MRS* may be deduced from (PROP.7): lower limit for *CRS*, and from (PROP.8): *MRS* is more indulgent than *CRS*.

$$MRS = \frac{a}{a+c},\ MPS = \frac{a}{a+b} \text{ and } |E| - 1 = a + b + c + d$$

The MUC scores are first computed and after that the resulting fractions are equated with those above (numerator *and* denominator), and the values of *a, b, c* and *d* are found. The κ coefficient is then computed using the definition below (Krippendorff 1980). The terms p_{Ah} and p_{Ao} represent the probability of a random agreement (on a link) and the proportion of agreements (out of all the possible links). They are eventually computed using the MUC scores.

(DEF.7)
$$\kappa = \frac{p_{Ao} - p_{Ah}}{1 - p_{Ah}} \quad \text{where } p_{Ao} = \frac{a+d}{a+b+c+d}$$
$$\text{and } p_{Ah} = \frac{(a+c)\cdot(a+b)+(c+d)\cdot(b+d)}{(a+b+c+d)^2}$$

The κ factor measures the agreement level above random, varying between –1 and 1: 1 means perfect agreement, 0 is the chance level (statistical independence) and –1 means perfect contrary correlation. On our sample text (cf. 1.4) we find $\kappa = -0.18$, that is, rather contrary correlation. Of course, inter-annotator agreement reaches values closer to 1.

There are three problems with this measure: (1) it uses (at least theoretically) coreference links, which have been shown to be less relevant than the sets; (2) replacing recall and precision by a single value is less informative; (3) κ is computed directly from *MRS* and *MPS*, so it seems unable to be more informative, even if it is less indulgent.

3.4 The B^3 measure (A. Bagga and B. Baldwin)

Starting from a similar observation of the MUC measure's indulgence, this measure attempts to penalize responses that amalgamate large RE classes – a sign that the system may use a trivial strategy. The scores are first computed for each RE: B^3 recall for a given RE of a key class *K* is the percentage of *K* that is contained in the response class *R* containing the given RE. Precision is computed symmetrically.

(DEF.8)
$$BRS(ER_i) = \frac{|R \cap K|}{|K|} \text{ and } BPS(ER_i) = \frac{|R \cap K|}{|R|}$$
where $ER_i \in R$ and $ER_i \in K$

For our sample text (cf. 1.4 and *Figure 2*), $BRS(ER_1)=2/2=1$ and $BRS(ER_3)=3/3=1$, this being true for all REs in K_1 and K_2. Then, $BRS(ER_6)=5/7$ and $BRS(ER_{11})=2/7$, these being the two possible values for the REs in K_3, and finally $BRS(ER_{13})=4/5$ and $BRS(ER_{17})=1/5$. To find the global recall and precision, the authors consider the average scores over all the REs, with two variants: either the REs or their classes have the same weight. No formula is given, but the authors seem to privilege the first option, hence:

(DEF.9)
$$BRS = \frac{1}{|E|} \cdot \sum_{\substack{K \in P_K \\ R \in P_R}} \frac{|R \cap K|^2}{|K|} \text{ and } BPS = \frac{1}{|E|} \cdot \sum_{\substack{K \in P_K \\ R \in P_R}} \frac{|R \cap K|^2}{|R|}$$

If all the classes receive the same weight in the average, the result is:

$$\text{(DEF.10)} \quad BRS' = \frac{1}{|E|} \cdot \sum_{\substack{K \in P_K \\ R \in P_R}} \left(\frac{|R \cap K|}{|K|} \right)^2 \text{ and } BPS' = \frac{1}{|E|} \cdot \sum_{\substack{K \in P_K \\ R \in P_R}} \left(\frac{|R \cap K|}{|R|} \right)^2$$

It is easy to prove that the upper limit (100%) is reached only for $P_R = P_K$, so that the upper limit criterion (1) is satisfied. However, 0% scores are never reached, as the scores for each RE are never null, so the lower scores criterion (5) is not satisfied – hence nor (4), nor (2). We can even prove the following inequality (cf. Appendix), showing that the B^3 measure is quite questionable in the low scores domain.

$$\text{(PROP.6)} \quad \frac{|P_K|}{|E|} \le BRS \le 1 \text{ and } \frac{|P_R|}{|E|} \le BPS \le 1$$

3.5 Core equivalence classes – C measure[6]

This measure is based on the concept of core equivalence classes: the core class $c^*(K)$ of a key class K is the response class that "best matches" K, i.e. the response class that contains most of K's REs. *All the REs* from a class K that are not in its core $c^*(K)$ count as recall errors, which is a less indulgent count than the MUC measure. To compute precision, we use the core class $c^*(R)$ of each response class R.

3.5.1 Example

Using the projections in *Figure 2*, we notice that K_1 and K_2 are included respectively in R_1 and R_2, so their core classes are $c^*(K_1) = R_1$ and $c^*(K_2) = R_2$. The largest projection of K_3 on P_R is on R_1 (R_1 includes five elements of K_3), so $c^*(K_3) = R_1$ and finally $c^*(K_4) = R_2$. The core response classes of K_1 and K_3, as well as K_2 and K_4, are identical, which reflects the observation that these key classes are not correctly differentiated in the response. Now, the largest among the projections of R_2 on P_K (shaded areas) is on K_4, so $c^*(R_2) = K_4$; notice that the core fragment $c(R_2) = c^*(R_2) \cap R_2 = \{RE_{13}, RE_{14}, RE_{15}, RE_{16}\}$. Also, $c^*(R_1) = K_3$ and $c^*(R_3) = K_4$.

As for recall errors, there are none for K_1 and K_2, only one for K_4 (RE_{17} outside its core class R_2) and two for K_3 (RE_{11} and RE_{12} outside the core class R_1) – the MUC measure counts for K_3 only one error. There are thus three errors out of 13 possible errors (K_1: 1, K_2: 2, K_3 : 6, K_4 : 4) hence the core recall success $CRS = 10/13 \approx 77\%$ – whereas $MRS = 11/13 \approx 85\%$.

A symmetrical computation yields the precision score. Indeed, the number of REs outside the core of a response class corresponds to "wrong links" in the response. There are two precision errors for R_1 (RE_1 and ER_2 outside the core class $c^*(R_1) = K_3$) five for R_2 (ER_3, ER_4, ER_5, ER_{11} and ER_{12} outside the core class K_4) and none for R_3. There are thus seven errors out of 14 possible (6 + 8 + 0) hence the core precision success $CPS = 7/14 \approx 50\%$ – whereas the MUC score is once again higher, $MPS = 11/14 \approx 79\%$.

[6] This measure was first described in (Popescu-Belis and Robba 1998b) and was designed for the system built by the author and I. Robba (LIMSI-CNRS, Orsay, France).

3.5.2 Definitions

We first define the sub-cores $c(K_i)$ and $c(R_j)$, that is, the largest fragment among the projections:

(DEF.11) $c(K) = \underset{A \in \pi(K)}{ArgMax} |A|$ and $c(R) = \underset{B \in \sigma(R)}{ArgMax} |B|$

When several fragments have the maximal size, one is chosen at random. The core classes are then defined as follows:

(DEF.12) $c^*(K) = R$ with $R \supset c(K)$ and $R \in P_R$
$c^*(R) = K$ with $K \supset c(R)$ and $K \in P_K$

Remember that the core class of a key class is a *response class* and the core class of a response class is a key class. The scores are symmetrical, and their formal expression is:

(DEF.13) $$CRS = \frac{\left(\sum_{K \in P_K} |c(K)|\right) - |P_K|}{|E| - |P_K|} \quad \text{and} \quad CRS = 1 \text{ if } |P_K| = |E|$$

(DEF.14) $$CPS = \frac{\left(\sum_{R \in P_R} |c(R)|\right) - |P_R|}{|E| - |P_R|} \quad \text{and} \quad CPS = 1 \text{ if } |P_R| = |E|$$

Choosing by convention that $CPS = 100\%$ when $|P_R| = |E|$ (no resolution) acknowledges the fact that there is certainly no wrong link. The convention $CRS = 100\%$ applies when $|P_K| = |E|$, i.e. there is no coreference in the key, so there is no possible recall error.

3.5.3 Properties

The C measure obviously satisfies the upper limit criterion (the case $\forall i$, $|c(K_i)| = |K_i|$ and the same for R_j). As for the lower limit criteria, the low scores criterion is satisfied, as the case when there is no resolution obtains zero recall, hence zero *f-measure*, except if the key is such as $|P_K| = |E|$ (nothing to solve). We have proved (cf. Appendix) the following inequalities:

(PROP.7) $$CRS \geq \frac{|R_m| - |P_K|}{|E| - |P_K|} \quad \text{and} \quad CPS \geq \frac{|K_m| - |P_R|}{|E| - |P_R|}$$
where K_m and R_m are the largest key and response classes

There are thus cases in which the direct lower limit criterion (4) is not satisfied. If the largest key class K_m is indeed very large, then a response with very few classes (e.g., $P_R = \{E\}$, total grouping) obtains a precision score above zero. This phenomenon is, however, less important as with the MUC measure, as the C measure is always more severe than the MUC measure, which was one of our goals (cf. also the examples in §5).

(PROP.8) • For a fixed RE set and partitions, $CPS \leq MPS$ and $CRS \leq MRS$
• $CRS = MRS \Leftrightarrow \forall K \in P_K$, $\pi(K) \setminus \{c(K)\}$ is a set of singletons
• $CPS = MPS \Leftrightarrow \forall R \in P_R$, $\sigma(R) \setminus \{c(R)\}$ is a set of singletons

3.5.4 An alternative: exclusive core classes – XC measure

Core classes attempt to grasp "the system's idea" about the correct classes. However, the definition does not state that all core classes should be distinct. We thus designed an algorithm to build *exclusive core classes xc*(K)* that are always distinct ($K_i \neq K_j \Rightarrow xc^*(K_i) \neq xc^*(K_j)$), so that confusion of two core classes (c^*) is penalized. The algorithm starts with the largest key class, and assigns exclusive core classes sequentially; once such a class (a response class) has been assigned, it is no longer available[7]. The symmetrical construction for the response classes is not meaningful here. For each $xc^*(K)$, the number of correct REs (i.e. from K) is the recall success, and the number of incorrect REs is the precision error (examples in §5.1).

Unfortunately, the algorithmic definition yields no simple formulae for this measure. Only the upper limit criterion is easy to verify. The experimental results (§5) show that this measure, which was intended to be more severe than the core measure, does not always fulfill this goal.

3.6 Transmission of referring information – H measure

We will briefly introduce here an application of information theory to reference understanding, i.e., constant co-activation of the same MRs in the sender and the receiver (cf. §1.2)[8]. This phenomenon is also found in information theory models of communication channels (Shannon and Weaver 1949), more specifically in the study of their capacity (Ash 1965).

In the communication channel model, the sender or source is a random variable that may take several values (here, the MR activated for each referring act), and the receiver or receptor is another random variable, with values from another set (the MRs activated on reception). The capacity of the channel (its accuracy or noiselessness) is measured using the statistical correlation of these two variables. The average emitted information per transmission (or here, referring act) is the *entropy* of the source, computed here using the number of REs in each key class. Accordingly, there is also an average received information. We define here the average *referring information* per referring act, for the sender and for the receiver:

(DEF.15) $$H(P_K) = -\sum_{K_i \in P_K} \frac{|K_i|}{|E|} \cdot \log \frac{|K_i|}{|E|} \; ; \; H(P_R) = -\sum_{R_j \in P_R} \frac{|R_j|}{|E|} \cdot \log \frac{|R_j|}{|E|}$$

An accurate communication channel guarantees maximal correlation between the two random variables, sender vs. receiver. The loss in the channel is defined as the conditioned entropy of the sender given the receiver value, averaged over these possible values. The conditioned entropy is computed using the probability law of the couple <sender variable, receiver variable>. Here, this law is given exactly by the set of all intersections of key classes

[7] On the sample text, the algorithm first assigns $xc^*(K_3) = R_1$, then $xc^*(K_4) = R_2$, then $xc^*(K_2) = \emptyset$ (as R_2 is no longer available) and $xc^*(K_1) = \emptyset$.

[8] This model is developed in (Popescu-Belis 1999b).

with response classes. We thus define the *loss of referring information* as the conditioned entropy of the sender given the receiver, noted $H(P_K|P_R)$. This quantity represents how much information about the sender (key RE sets) is lost for the receiver (response RE sets).

Going beyond information theory, we also define the *unjustified accrual of referring information* as the conditioned entropy of the receiver given the sender, noted $H(P_R|P_K)$. This quantity accounts for an increase in the receiver's referring information without any relevance or justification from the sender. The exact definitions are:

(DEF.16)
$$H(P_K|P_R) = -\sum_{(K_i,R_j)\in P_K\times P_R} \frac{|K_i \cap R_j|}{|E|} \cdot \log \frac{|K_i \cap R_j|}{|R_j|}$$
$$H(P_R|P_K) = -\sum_{(K_i,R_j)\in P_K\times P_R} \frac{|K_i \cap R_j|}{|E|} \cdot \log \frac{|K_i \cap R_j|}{|K_i|}$$
(with the convention that "$0\cdot\log(0) = 0$")

The following properties of conditioned entropy (Ash 1965) indicate that our interpretation of these notions is coherent. Indeed, the first line in (PROP.9) is in our view the *fundamental equation of referring information*, as it reads: "the received referring information equals the information sent, minus the losses, plus the unjustified accruals".

(PROP.9)
- $H(P_R) = H(P_K) - H(P_K|P_R) + H(P_R|P_K)$
- $0 \le H(P_R|P_K) \le H(P_R)$
- $0 \le H(P_K|P_R) \le H(P_K)$

It is quite natural to define precision errors as information loss, and recall errors as unjustified information accrual, thus defining an entropy-based measure (H). The inequalities in (PROP.9) insure that these values are positive and lower than the information encoded by the sender or the receiver. Thus:

(DEF.17)
$$HRS = \frac{H(P_R) - H(P_R|P_K)}{H(P_R)} \quad \text{and} \quad HPS = \frac{H(P_K) - H(P_K|P_R)}{H(P_K)}$$
with $HRS = 1$ if $H(P_R) = 0$, and $HPS = 1$ if $H(P_K) = 0$

A non-trivial result from information theory (Ash 1965) allows us to prove that the H measure satisfies the upper limit criterion (1). Indeed:

(PROP.10) *f-measure* = 100% $\Leftrightarrow$ $H(P_R|P_K) = H(P_K|P_R) = 0 \Leftrightarrow P_R = P_K$

Quite nicely, it is possible to characterize precisely which responses yield a 0% score (*f-measure*), which proves that the H measure satisfies the lower limit criteria (2-5). Of course, evaluators still have to agree that the responses below are indeed the worst possible.

(PROP.11) *f-measure* = 0 iff at least one of the following conditions holds:
- $H(P_R) = 0$ & $H(P_K) \neq 0$ (one resp. class, several key classes)

- $H(P_k) = 0$ & $H(P_R) \neq 0$ (one key class, several resp. classes)
- $H(P_k) \neq 0$ & $H(P_R) \neq 0$ & "P_K and P_R are independent"

The last condition is the one in which knowing the response partition of REs brings no knowledge about the key partition (statistical independence), and it can be characterized as follows.

(PROP.12) The following conditions are equivalent:
- "P_K and P_R are independent"
- $H(P_K) = H(P_K|P_R)$
- $H(P_R) = H(P_R|P_K)$
- vectors $(|K_1 \cap R_j|, \ldots, |K_n \cap R_j|)$, $1 \leq j \leq m$, are proportional
- vectors $(|K_i \cap R_1|, \ldots, |K_i \cap R_m|)$, $1 \leq i \leq n$, are proportional

In other words, each key class K_i must project onto P_R following the same proportions – and that is equivalent to the reciprocal condition. This shows that for a given key, and for a given number of response classes $|P_R|$, it is not always possible to reach a 0% score, as the key classes K_i contain an integer number of REs, and are not freely dividable. Of course, if $|P_R|$ is not fixed, then $|P_R| = 1$ will do the job, but is not always possible to attain 0% with a non trivial response, i.e. $|P_R| > 1$. Thanks to the H measure's strong theoretic background, its lower limits are easier to analyze. Besides, this measure does not seem constantly more severe or more indulgent than the other measures, as the numeric results will show.

3.7 Coarse evaluation using distributional measures

Comparing the number of key and response classes offers a simple way to estimate the quality of the response. Roughly speaking, if there are much more response classes than key classes, then probably recall errors outnumber precision errors, and the other way round. Of course, having the same number of classes does not guarantee *at all* that key and response coincide.

A slightly more complex idea is to compare also the sizes of the classes after sorting them and completing the series. For instance, in our example text, the key sizes are (7, 5, 3, 2) and the response sizes (9, 7, 1, '0'). A distance between these vectors is thus |9–7| + |7–5| + |1–3| + |0–2| = 8, with the maximum distance being 32 (between (17, 0, …, 0) and (1, …, 1)). The *distributional match* (*DMT*) between key and response is thus 1–(8/32) = 75%. The general formula is:

(DEF.18) $DMT = 1 - \frac{1}{|E|} \cdot \sum_i \left| |K_i| - |R_i| \right|$ where K_i and R_i are rearranged so that $|K_1| \geq |K_2| \geq \ldots \geq |K_n|$ and $|R_1| \geq |R_2| \geq \ldots \geq |R_m|$,
and $K_i = \varnothing$ if $i \geq |P_k|$, and $R_i = \varnothing$ if $i \geq |P_R|$

It would also be interesting to know whether the largest classes occur in the key or in the response, but the absolute values in the *DMT* measure make it symmetrical with respect to P_K and P_R. On our sample text, however, R_1 and R_2 are larger than K_1 and K_2, but R_3 and 'R_4' are smaller than K_3 and K_4; so, the response has put together too many REs. The representation *Figure 4* of an example shows the main idea behind the *distributional error* (*D-err*) measure:

order the K_i and R_j according to size, then consider on one side the positions where $|K_i| \geq |R_i|$ (upper white zones) and on the other side the positions where $|K_i| < |R_i|$ (upper dark gray zones). Then, compute the average position of the first group vs. the second: if it is smaller, this means that on average the response classes are smaller than the key classes (and also more numerous). This yields a *D-err recall error*, and in the opposite case a *D-err precision error*, indicating the main flaw in the response. The exact formulae are a bit tedious.

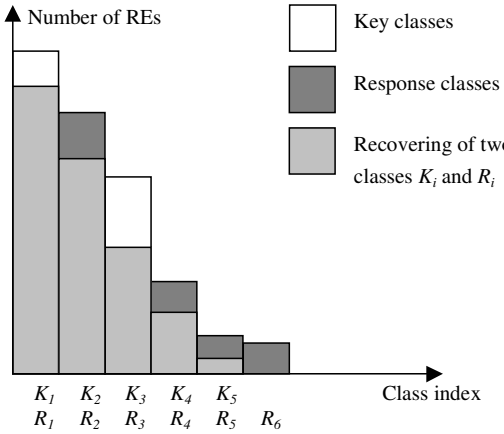

Figure 4. Comparative sizes of key and responses classes (sorted)

The *DMT* and the *D-err* scores vary between 0% and 100% (*D-err* also indicates 'recall' or 'precision'), but a score of 100% does not mean that the response structurally matches the key. Even if they do not satisfy the upper limit criterion (1), they do satisfy the direct lower limit criterion (2): low distributional scores signal poor responses. These measures should not be used alone, but they give an overall idea of a system's main bias.

4. GENERALIZATIONS

4.1 Evaluation with respect to an elementary strategy

R. Mitkov (1998) estimates the improvements of an anaphora resolution system against a simplistic strategy or *baseline*. This evaluation method supposes in fact that an evaluation measure has already been chosen and makes a *differential* use of it, by fixing the 0% score at the level of a simplistic strategy. The new measure automatically satisfies the low scores criterion (5).

Any of the previous measures could be chosen here, but what about the baseline? As grouping all the REs into one response class yields 100% recall, and grouping no or few REs yields a good precision score, these seem too extreme to serve as a baseline. Another idea is to define the baseline strategy according to the knowledge it requires, e.g., use only of the number and gender of the REs, or of their first noun. Yet another idea is to compare the system's response to a random response – e.g., a random partition of the REs that has the same distributional profile as the key.

We have used differential evaluation to estimate the contribution of each piece of knowledge to our system's results (Popescu-Belis and Robba 1998a). Some of the rules have been

alternatively deactivated, then pairs of rules have been deactivated, and the scores compared, showing that the most important rule was semantic agreement. To increase the reliability of the scores we averaged several evaluation measures.

4.2 The status of singleton classes

If coreference links are to be used, the REs that are not linked to others in the key should not be considered for evaluation. Indeed, the MUC-6 summaries mention only classes with two or more REs, despite the fact that the MUC measure does not exclude singletons. The MUC measure, as well as others, actually must count singletons in the response, among the projections of the key classes.

We compare two manners of calculating the MUC recall, first with all $K \in P_K$, then only with classes having two or more elements (K such as $|K| \geq 2$, say $K \in P^2_K$). We do the same for MUC precision, and show that for the MUC and for the core-class measure, singletons do not influence the result:

(PROP.13)
- Recall scores *MRS* and *CRS* are invariable whether they are computed using P_K or $P^2_K = \{K \mid K \in P_K \wedge |K| \geq 2\}$
- Precision scores *MPS* and *CPS* are invariable whether they are computed using P_R or $P^2_R = \{R \mid R \in P_R \wedge |R| \geq 2\}$

Such results do not hold for the B^3, the XC or the H measures. For the latter, the entropy of the sender or that of the receiver clearly depends on the singletons too. The κ measure makes use of both *MRS* and *MPS* so singletons have to be considered (in P_R for *MRS* and in P_K for *MPS*). As a conclusion, it is more homogenous and coherent to always consider the singletons when computing recall and precision.

4.3 Different RE sets for the key and the response

We have until now defended the modularity of evaluation (cf. §2.1), according to which reference resolution should be evaluated using only the correct set of REs . Identification of REs should be evaluated using separate recall and precision measures (the program misses correct REs vs. finds wrong ones). However, if the identification of REs and their correct resolution are evaluated together, then the key P_K and the response P_R are no longer partitions of the same RE set E, and the preceding measures must be adapted.

For instance, the description of the MUC measure (1995) makes use of identical RE sets for P_K and P_R, but what was implemented for MUC-6 and 7 also works when this is not the case (the implementation actually fits our definition below (DEF.19)). The authors of the B^3 measure do not discuss this issue as their measure is designed for inter-document coreference and supposes that the system knows the correct entities for each document. Our proposal here is straightforward:

(DEF.19) If the key and response RE sets are different ($E_K \neq E_R$) then:
- let $E = E_K \cup E_R$,
- let $P'_K = P_K \cup \{ \{re\} \mid re \in E_R \setminus E_K \}$ ("add singletons to P_K")
- let $P'_R = P_R \cup \{ \{re\} \mid re \in E_K \setminus E_R \}$ ("add singletons to P_R")
- use the above measures with E, P'_K and P'_R.

Using E, P'_K and P'_R does not affect the MUC and core-class measures (PROP.13) as they only add singletons.

4.4 Restriction to an RE subset. Anaphora resolution

In case reference resolution has to be evaluated for a particular subset of REs, for instance proper names, the simplest solution is to restrict the RE set E as well as the key and response partitions P_K and P_R to this subset, discarding all other REs (but possibly also valuable links). Another idea is to compute recall and precision scores for each RE using the B^3 measure, and use only the relevant RE subset for the final average[9].

This strategy does not apply to anaphora resolution, which *is not* the restriction of reference resolution to pronouns. As anaphora resolution requires the attachment of pronouns to non-pronominal antecedent REs (i.e., more than pronoun grouping), its evaluation requires the total set of key classes: there is no "unique" antecedent for a given pronoun, but a whole class. So, as criticisms from pragmatics have suggested (e.g., (Reboul 1994)), pronouns should be considered as full REs, and the RE–MR links privileged over the pronoun–antecedent links.

An attempt to evaluate anaphora resolution from within the MR paradigm is the following[10]: for each pronominal RE, see if its response class contains at least one non-pronominal RE from its key class, and if it does not, count a recall error. This is too indulgent, as it is not enough for a pronoun and an "antecedent" to be in the same response class, if this contains also a lot of wrong "antecedents" – these are then precision errors for the respective pronouns. This is however too severe, as our experiments showed.

Another evaluation option is based on responses made of pronoun–antecedent links, but requires the full key classes (all REs). If a <pronoun, antecedent> couple of the response does not belong to the same key class, then this an "error", otherwise a "success". Following for instance (Mitkov and Belguith 1998), "recall" could be defined as the success rate compared to the number of pronouns, and "precision" could be defined as the success rate compared to the number of pronouns processed by the program. However, one may wonder whether these definitions preserve the typical meaning of "recall" and "precision".

4.5 Various types of coreference

From a linguistic point of view, it has been noted that referring relations among REs also include relations such as whole/part, type/token, individual/function, variable/value, etc. Proposals have been made for a specific annotation of such relations between REs (Bruneseaux 1998). In order to evaluate the understanding of these relations, the MUC-7 guidelines (Hirschman 1997) assimilated some of them to standard coreference at a given point in the text (such as an individual and their function at the mentioned moment), while discarding the others. Standard measures may then be used.

These relations, however, are not strictly speaking relations between REs, but between mental representations (Popescu-Belis, et al. 1998), for instance the MR for an individual and the MR for a function or job. Hence, there are two levels to evaluate, starting with the one that is extensively analyzed in this paper, namely the correct activation of MRs upon RE

[9] It has to be noticed that B^3 recall and precision per RE coincide for all the REs in a given $K_i \cap R_j$, so they may not characterize a *specific* RE.

[10] This was independently proposed by S. Azzam *et al.* (1998), and by I.Robba and the author.

sending/reception (these are the "identity reference" links between REs). Then each type of referring link between MRs should be evaluated separately, using the correct MR set and the key/response partitions for this set according to each referring phenomenon. The specificity of each phenomenon requires further analysis that is beyond our present scope.

5. EXAMPLES

5.1 Synthetic keys and responses

We first use some artificial examples to illustrate the measures on particular cases. We created artificial texts annotating the REs, then used our system to enter a key and a response, and automatically trigger the measures. We tested the following examples:

1. The sample text (plus key and response) from Section 1.4.
2. A text with ten REs and two key classes, $K_1 = \{1, 2, 3, 4, 5\}$ and $K_2 = \{6, 7, 8, 9, 10\}$. The response is first a "no resolution" response, i.e. $P_R = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10\}\}$.
3. Using the same text, we suppose now that the response groups all REs into one class, $R_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, hence $P_R = \{R_1\}$.
4. The sample text from the MUC-6 proceedings. There are 147 REs in 15 key classes (singletons are not counted). First, we consider the "no resolution" response, with $|P_R| = 147$.
5. Using the same text, we suppose now that the response groups all REs into one class, so that $|P_R| = 1$.
6. The same text contains in fact 50 pronouns, but only five key classes contain pronouns. We suppose now that the system is unable to understand pronouns, so it groups into one separate response class, but correctly solves the 97 other REs. So, $|P_R| = 15 + 1$.

Table 3. Recall and precision for the sample texts (%)

	MRS	*MPS*	*BRS*	*BPS*	*CRS*	*CPS*	*XRS*	*XPS*	*D-err*	*DMT*	*HRS*	*HPS*
1	85	79	74	49	77	50	53	59	34-p	75	55	37
2	0	100	20	100	0	100	20	100	41-r	11	30	100
3	100	89	100	50	100	44	50	50	51-p	44	100	0
4	0	100	10	100	0	100	10	100	39-r	10	40	100
5	100	90	100	19	100	31	31	31	58-p	31	100	0
6	96	97	65	79	67	82	69	84	7-r	86	76	81

Table 4. F-measures (in %) and kappa for the designed examples

Example	MUC	B^3	κ	C	XC	H
1	81	59	–18	61	56	44
2	0	33	0	0	33	46
3	94	67	0	62	50	0
4	0	19	0	0	19	57
5	95	33	0	47	31	0
6	97	71	66	73	76	78

Results are given in *Table 3* (recall and precision percentage, two digits) and *Table 4* (*f-measures* percentage, two digits). Example (1) obtains relatively high scores (except for κ) despite a confuse response (cf. Section 1.4). Examples (2) and (4) show the scores of a system that "performs no resolution", whereas examples (3) and (5) are those of a system that "groups

all REs". Both strategies are extremely crude, but we see they sometimes receive high scores, beside the expected fact that precision is 100% in the first case, and recall is 100% in the second case (except for the XC measure). The *f-measures* in examples (3) and (5) are quite high, especially for the MUC measure, which proves to be too indulgent. Example (6) is still more realistic, and shows again the indulgence of some measures, in a case when 30% of the REs (the pronouns) are incorrectly resolved.

Regarding relative indulgence, these results confirm that the only comparable measures are C and MUC, the former being less indulgent. For any other pair of measures, there is no constant relationship over all the examples (except, here, the κ measure). These scores are also covariant, that is, increase or decrease simultaneously between two examples. For instance, (6) receives better *f-measure* scores than (4) and (5), for all the measures.

5.2 Results of our system on real texts

We have developed a reference processing workbench (Popescu-Belis and Robba 1998a, Popescu-Belis *et al.* 1998) which integrates all the above measures. An annotation module provides the user an interface to annotate key REs and key classes, and converters to/from various SGML annotation formats (Bruneseaux 1998, Popescu-Belis 1998).

The main program is the reference resolution module for texts in French, which constructs response RE classes. Its algorithm (Popescu-Belis *et al.* 1998) parallels the one proposed by (Lappin and Leass 1994). The REs are processed one by one, and for each RE the program either activates an existing MR or creates a new one, in both cases the current RE being added to the activated MR. For each RE, the program thus determines the set of MRs that are candidates for activation, by computing an average compatibility between the current RE and each MR, or more exactly the REs that constitute it. The compatibility between two REs depends on their gender (in French), number and semantic content (head and determiners of the noun group). Among candidate MRs, the most salient one is activated and its salience is updated; if there is none, a new MR is created.

Table 5. Numeric characteristics of the trial texts

Characteristic	VA	LPG.eq	LPG
Words	2630	7405	28576
REs ($\|E\|$)	638	686	3359
Key MRs ($\|P_K\|$)	372	216	480
Coref. rate ($\|E\| / \|P_K\|$)	1.72	3.18	7.00
Noun phrase REs	510	390	1864
Pronoun REs	102	262	1398
Non parsed REs	26	34	97

Experiments with real texts are made difficult by the necessity to define the key for potentially long texts (Bruneseaux 1998, Popescu-Belis 1998). We have used a short story by Stendhal (annotated at LIMSI-CNRS, Orsay, France) and a fragment of a novel by Balzac (annotated at LORIA, Nancy, France), both 19th century French authors. The first is noted VA, the second LPG, and LPG.eq is a fragment of LPG with as many REs as VA (cf. *Table 5*). The texts are quite long (ca. 100 pages for LPG) and have important coreference rates ($|E| / |P_K|$).

Table 6. System's results on trial texts (in %)

MRS	*MPS*	*BRS*	*BPS*	*CRS*	*CPS*	*XRS*	*XPS*	*D-err*	*DMT*	*HRS*	*HPS*

	MRS	*MPS*	*BRS*	*BPS*	*CRS*	*CPS*	*XRS*	*XPS*	*D-err*	*DMT*	*HRS*	*HPS*
VA	70	78	75	75	53	47	70	79	15-R	85	89	89
L.eq	62	77	50	57	43	36	41	65	17-R	73	71	71
LPG	70	88	37	52	43	44	35	61	14-R	66	59	64

Table 7. F-measures (in %) and κ for the system's results on trial texts (from *Table 6*)

	MUC	B^3	κ	C	XC	H
VA	74	75	57	50	74	89
LPG.eq	69	53	20	39	50	71
LPG	78	43	9	43	44	61

Our program's results (cf. *Table 6* and *Table 7*) may seem quite high when compared to programs from the MUC campaign, which scored in the 60% range. In fact, there are two differences in evaluation: we do not evaluate RE identification, giving the system the correct REs from the start, and we use much longer texts with larger key classes, thus biasing the MUC measure towards higher scores, as noted in §3.2.

Despite the similar nature of the three texts, the scores of the program are quite variable. Indeed, the MUC measure, and to a lesser degree the C measure, are more indulgent as the number of REs increases (VA vs. LPG and LPG.eq vs. LPG), while the "system's quality" is constant. The B^3, XC and H measures vary in the opposite direction. So, *f-measures* increase for MUC and C, and decrease for B^3, XC and H when comparing LPG.eq and LPG, because they do not have the same bias with respect to the number of REs. However, when applied to texts of similar lengths, all the measures agree in designating the response on VA as better than the one on LPG.eq, reflecting the capacities of the program on those texts.

6. CONCLUSION

In this paper, we have introduced a framework for reference transmission and another one for system evaluation, which we hope are easy to agree upon, in their broad lines. Using a small number of introductory definitions (the main one being the projection of a class), we have provided precise and unified formulae for the already proposed measures of reference understanding. In order to answer some of the inadequacies of these measures, we proposed two new measures, the core-class and the information-based measures, and established some of their properties. We also discussed two possible extensions, the exclusive-core-class and the distributional measures, and provided various numerical results for all measures.

We might now conclude that none of these measures seems to prevail due to intrinsic qualities. The information-based measure, though, proceeds from a strong theoretical background. It may be suggested that each measure may be suited to a certain style of input data, or to a certain quality level of the programs. In addition, despite the fact that no measure seems able to grasp the very nature of reference understanding, it is likely that the preceding measures, elaborated by different authors, provide different views of the quality of a program. If *all these measures are unanimous* in declaring a response better than another, then it is legitimate to consider that this response *really is better* than the other.

REFERENCES

Ash Robert B. 1965, *Information Theory*, Interscience Publishers (John Wiley and Sons), New York, NY.

Azzam Saliha, Kevin Humphreys and Robert Gaizauskas 1998, Evaluating a Focus-Based Approach to Anaphora Resolution, Proceedings COLING-ACL '98, Université de Montréal, Montréal, Québec, Canada, volume I/II, p. 74-78.

Bagga Amit and Breck Baldwin 1998a, Algorithms for Scoring Coreference Chains, Proceedings LREC'98 Workshop on Linguistic Coreference, Granada, Spain.

Bagga Amit and Breck Baldwin 1998b, Entity-Based Cross-Document Coreferencing Using the Vector Space Model, Proceedings COLING-ACL '98, Université de Montréal, Montréal, Québec, Canada, volume I/II, p. 79-85.

Bruneseaux Florence 1998, Noms propres, syntagmes nominaux, expressions référentielles, *Langues : cahiers d'études et de recherches francophones, 1*, 1, p. 46-59.

Grishman Ralph and Beth Sundheim 1996, Message Understanding Conference-6: A Brief History, Proceedings 16th International Conference on Computational Linguistics (COLING-96), Center for Sprogteknologi, Copenhagen, p. 466-471.

Hirschman Lynette 1997, MUC-7 Coreference Task Definition 3.0, MITRE Corp.

Hirschman Lynette 1998, Language Understanding Evaluations: Lessons Learned from MUC and ATIS, Proceedings First International Conference on Language Resources and Evaluation (LREC '98), ELRA, Granada, Spain, volume 1/2, p. 117-122.

Krippendorff Klaus 1980, *Content Analysis: An Introduction to Its Methodology*, Sage Publications, Beverly Hills, CA.

Lappin Shalom and Herbert J. Leass 1994, An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics, 20*, 4, p. 535-561.

Mitkov Ruslan 1998, Robust pronoun resolution with limited knowledge, Proceedings COLING-ACL '98, Université de Montréal, Montréal, Québec, Canada, volume II/II, p. 869-875.

Mitkov Ruslan and Lamia Belguith 1998, Pronoun resolution made simple: a robust, knowledge-poor approach in action, Proceedings TALN '98, Paris, p. 42-51.

MUC-6 1995, *Proceedings of the 6th Message Understanding Conference (DARPA MUC-6 '95)*, Morgan Kaufman, San Francisco, CA.

Passonneau Rebecca J. 1997, Applying Reliability Metrics to Co-Reference Annotation, Technical Report Columbia University - Department of Computer Science, CUCS-017-97.

Popescu-Belis Andrei 1998, How Corpora with Annotated Coreference Links Improve Anaphora and Reference Resolution, Proceedings First International Conference on Language Resources and Evaluation (LREC'98), ELRA, Grenade, Espagne, volume 1/2, p. 567-572.

Popescu-Belis Andrei 1999a, L'évaluation en génie linguistique : un modèle pour vérifier la cohérence des mesures, *Langues (Cahiers d'études et de recherches francophones), 2*, 2, p. 151-162.

Popescu-Belis Andrei 1999b, *Modélisation multi-agent des échanges langagiers : application au problème de la référence et son évaluation*, Thèse d'université, Université de Paris XI (Paris-Sud).

Popescu-Belis Andrei and Isabelle Robba 1998a, Evaluation of Coreference Rules on Complex Narrative Texts, Proceedings Second Colloquium on Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2), University Centre for Computer Corpus Research on Language, Lancaster, UK, p. 178-185.

Popescu-Belis Andrei and Isabelle Robba 1998b, Three New Methods for Evaluating Reference Resolution, Proceedings LREC'98 Workshop on Linguistic Coreference, Granada, Spain.

Popescu-Belis Andrei, Isabelle Robba and Gérard Sabah 1998, Reference Resolution Beyond Coreference: a Conceptual Frame and its Application, Proceedings COLING-ACL '98, Université de Montréal, Montréal, Québec, Canada, volume II/II, p. 1046-1052.

Reboul Anne 1994, L'anaphore pronominale : le problème de l'attribution des référents, *Langage et pertinence*, Presses Universitaires de Nancy, Nancy, p. 105-173.

Shannon Claude Elwood and Warren Weaver 1949, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Ill.

Van Rijsbergen Cornelis J. 1979, *Information Retrieval*, Butterworth, London.

Vilain Mark, John Burger, John Aberdeen, Dennis Connolly and Lynette Hirschman 1995, A Model-Theoretic Coreference Scoring Scheme, Proceedings 6th Message Understanding Conference (MUC-6), Columbia, MD, p. 45-52.

APPENDIX

The theorems whose demonstrations are not obvious from the text are sketchily demonstrated here.

(PROP.2) – In the *MRS* numerator, $|\pi(K_i)|$ can be written as $\Sigma_{R_j \in P_R} 1_{K_i \cap R_j}$, and in the *MPS* numerator, $|\sigma(R_j)|$ can be written as $\Sigma_{K_i \in P_K} 1_{K_i \cap R_j}$, where $1_{K \cap R}$ is 1 if $K \cap R \neq \varnothing$ and otherwise 0. The resulting expressions are the same for *MRS* and *MPS*.

(PROP.3) – The '$\Leftarrow$' sense is obvious: if each K is included in an R, then its projection is not fragmented, i.e. $|\pi(K)|=1$, so $MRS=1$. For the '$\Rightarrow$' sense, the error being a sum of positive values, each K must be such as $|\pi(K)|=1$, i.e. each K intersects only one response class, in which it is contained (q.e.d.). The proof is analogous for precision, and the result on the *f-measure* is a consequence of the first two.

(PROP.5) – The MRS numerator is a sum with $|P_K|$ terms, each of them smaller than $|P_R|$ because a key class K projects in at most $|P_R|$ fragments, one per response class. Analogous proof for MPS.

(PROP.6) – In the double sum from BRS, let us group the terms corresponding to the same key class K_i, and first show that $|K_i|^2 \geq \mathbf{\Sigma}_{R_j \in P_R} |K_i \cap R_j|^2 \geq |K_i|$. As $\mathbf{\Sigma}_{R_j \in P_R} |K_i \cap R_j| = |K_i|$, it is enough to square the terms to establish the first inequality; the second one is established using $|K_i \cap R_j|^2 \geq |K_i \cap R_j|$. To prove the result, the inequalities are divided by $|K_i|$, then summed on all K_i. Analogous proof for BPS.

(PROP.7) – To prove the result on CPS, let us pick a key class K_i. Then, $\forall R_j \in P_R$, $|c(R_j)| \geq |K_i \cap R_j|$ (because by definition the core fragment is the largest projection), so $\mathbf{\Sigma}_{R_j \in P_R} |c(R_j)| \geq \mathbf{\Sigma}_{R_j \in P_R} |K_i \cap R_j| = |K_i|$. Now, if we chose $K_i = K_m$, the largest key class, the result is established. Analogous proof for CRS.

(PROP.8) – If the denominator is zero, $MRS \geq CRS$ is true (cf. conventions). Otherwise, we have to prove the inequality of the numerators, which is easily written as $\mathbf{\Sigma}_{K_i \in P_K} (|K_i| - |\pi(K_i)|) \geq \mathbf{\Sigma}_{K_i \in P_K} (|c(K_i)| - 1)$. The meaning of the inequality $|K_i| \geq |c(K_i)| + |\pi(K_i)| - 1$ is that if we add to the REs in the core fragment of K_i one RE per other projection (not core), then we have less REs than in the whole K_i. This is obvious, and the equality happens iff all the projections of K_i, except maybe the core fragment, are singletons. That has to hold $\forall K_i \in P_K$ for $MRS = CRS$. Analogous proof for $MPS \geq CPS$.

(PROP.13) – For recall, we compare the formulae for E and P_K with those for E^2 and ${P^2}_K$ (i.e. where the singleton K_i and the corresponding REs have been removed). The denominator of MRS and CRS, that is $|E| - |P_K|$, doesn't change in this operation because the same number of REs is removed from E and from P_K to arrive at E^2 and ${P^2}_K$. We may rewrite the numerators as $\mathbf{\Sigma}_{K_i \in P_K} (|K_i| - |\pi(K_i)|)$ and $\mathbf{\Sigma}_{K_i \in P_K} (|c(K_i)| - 1)$, making obvious the fact the singletons do not count, as they have one projection and their core fragment is a singleton.