The Reference Annotation Framework: A case for semantic content representation

Susanne Salmon-Alt, Laurent Romary

The Reference Annotation Framework: A case for semantic content representation

Susanne Salmon-Alt*, Laurent Romary**

3rd December 2004

*ATILF - CNRS Nancy, France

** LORIA - INRIA Vandoeuvre-lès-Nancy, France

Susanne.Salmon-Alt@atilf.fr, Laurent.Romary@loria.fr

1 Introduction

This paper presents the Reference Annotation Framework (RAF), a specification platform for structural constraints on reference annotation schemes. By presenting the main components of the RAF meta-model as well as an overview of the data categories used to characterize them, we pursue two objectives: firstly, to show that RAF may function as a reading guide for existing and new task definitions covering a wide range of reference, coreference and anaphora annotation issues; secondly, to present a general methodology for the design of a whole class of semantic and pragmatic annotation schemes based on a simple markable-link paradigm. Indeed, the task of (co)reference and anaphora annotation basically consists in (a) the identification of the textual elements that may be considered as referring in a given discourse and (b) the kind of links those markables establish one with another. Both subtasks are strongly related to other issues in semantic content representation, such as named entities, informational status, discourse and dialogue structure, or semantic roles. Hence, our overall goal is to contribute to the establishments of standards in the domain of multi-modal semantic content representation [8], related to the activities of the ISO TC 37/SC 4.

In no case, RAF is intended to provide or to replace theoretically well founded task definitions for (co)reference annotation, as one can found in [11], [23] or [24]. On the contrary, RAF focuses on structural properties of a pivot representation format for reference related phenomena. The first underlying goal of RAF is to propose a representation format that is generic enough to be instantiated for the annotation of a whole range of phenomena considered in the literature to be coreferential or anaphoric. Those include generally coreference properly speaking [24] as well as so-called bridging relations [14], one-

and other-anaphora [18] and particular cases of pronominal reference such as pay-cheque and bound anaphora [2]. Other authors extend them to noun phrase predication [20], function-value relations [23], type identity [4], VP ellipsis [19], deictic and anaphoric reference to single and multi-clause units [12] [9], or tense [3]. This wide variety advocates strongly in favour of a definition of markables as autonomous linguistic objects, abstracting for example from text linearity or prior text segmentation. It also underlines the necessity of leaving open for concrete task definitions the precise features associated to the markables and the characterisation of the relationships to be encoded. Therefore, RAF provides a minimal set of mandatory core components (a meta-model), associated with descriptors (data categories) that may actually be instantiated in a modular and dynamic way, by making use of an externalized and centralized registry of data categories [16].

The second goal for RAF is to ensure that the definition and the articulation of the basic components of the meta-model allow to represent in an intuitive and consistent way structural properties of (co)reference annotations, such complex, embedded or disjoint markables (e.g. plural antecedents built conjointly upon two previous mentions), uncertainty (e.g. an annotator not being sure about his decision or a non deterministic anaphora resolution tool adding confidence scores to its output), disjoint or conjoint lectures of more than one annotated information (e.g. a given expression that refers back, in a mutual excluding way or not, to more than one markable).

2 Overview of the Reference Annotation Framework

The linguistic annotation framework (LAF) [17] provides a modelling approach to the specification of linguistic annotation schemes by identifying a class of structures that can be fully determined as the combination of two elements: a meta-model which is an abstract representation of the underlying structure shared by all models within a given field of linguistic representation; and a set of data categories, attached to the nodes of the meta-model, which bear the actual linguistic information of the annotation scheme. From the general LAF principles, it is possible to derive a meta-model that covers the main characteristics of reference related annotation schemes. Figure 1 outlines the proposal of such a meta-model. It is organized around three main components and gathers up all information related to a specific annotation document within a component named /Referential data collection/. Besides a /Global information/ component for the metadata associated with the annotation file, the specific linguistic objects of interest for any reference annotation project are /Referential markable/ and /Referential link/. Since referential markables are not systematically isomorphic to textual chunks identifiable at lower linguistics levels, they have to be understood a an abstraction over them. As a first approximation, referential markables may be seen as corresponding to the elements referred to in the lit-

erature as discourse entities [23], whereas referential links are the entities used to encode the relation holding between referential markables.

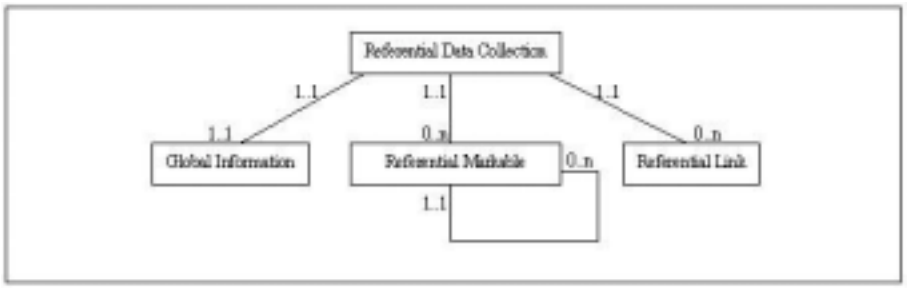

Figure 1: RAF meta-model

3 Referential markables

3.1 The meta-model component

The core task of any reference annotation procedure consists in establishing links between markables. Whether the identification of markables (e.g. anaphora and antecedents) has to be part of the reference annotation task properly speaking or not is matter of discussion [23] [24] and should be fixed in the task definition. In any case, RAF has to provide a representation for the linguistic objects to be linked, independently from their actual selection procedure (manually annotated from row data, automatically annotated based on the output of a POS-tagger or a parser) or their location (within the same text document or in an external source file). RAF provides such a representation in terms of referential markables. As the most important characteristic, referential markables are autonomous linguistic objects (i.e. represented by their own meta-model component. They have a mandatory feature /Source text/ that either contains the annotated source text (Figure 2) or points as a stand-off annotation to externalized source data, for instance words or syntactic chunks (Figure 3). In the latter case, one may optionally reproduce the text string for sake of legibility (Figure 4).

```
<struct id="m_1"  type="referentialMarkable">
      <feat type="sourceText">French Boss Aimé Jaquet</feat>
</struct>
praised
<struct id="m_2"  type="referentialMarkable">
      <feat type="sourceText">his</feat>
</struct>team's application.
```

Figure 2: In-line representation of markables

```
<w id="w_1">Frenchc</w>
<w id="w_2">boss</w>
<w id="w_3">Aimé</w>
<w id="w_4">Jacquetc</w>
<w id="w_5">praised</w>
<w id="w_9">his</w>
<w id="w_10">team's</w>
<w id="w_10">applicationc</w>
```

Figure 3: External source data file, segmented into words

```
<struct id="m_1"  type="referentialMarkable">
      <feat type="sourceText" target="w_1..w_4">French Boss Aimé Jaquetc</feat>
</struct>

<struct id="m_2"  type="referentialMarkable">
      <feat type="sourceText" target="w_9">his</feat>
</struct>
```

Figure 4: Stand-off representation of markables

The main argument in favour of referential markables as autonomous elements is the fact that they are not necessarily isomorphic to elements identifiable at lower linguistic levels, for instance words, syntactic chunks (nominal, pronominal or verbal phrases) or clause units. On the one hand, referential markables may be provided by parts of those units (incorporated pronouns: sp. *dámelo*, genitive markers: *his team's application*) or by missing elements (personal pronoun ellipses: it. *e' andato a mangiare*, case markers in Japanese). On the other hand, markables may also be built upon complex syntactic or discursive structures. This is frequently the case for clause-level antecedents of deictic demonstrative pronouns [12] and for antecedents of plural anaphora, occurring either as syntactically conjoint constituents (*John and Mary... they*), as enumerated items (*John, Mary and Sue... they*) or even as completely disjoint constituents (*Mary ... Bob ... they*). In those cases, the annotator should still have the possibility to conceive complex markables, i.e. to build them independently of whether the corresponding textual elements are actually contiguous or not. Figure 5 shows how RAF accounts for such a complex markable, built on non contiguous text chunks. Finally, considering markables as autonomous entities also allows for annotating referential expressions by cross-references to other data sources than text, like the original sound or video recordings for spoken or multimedia data. Here, the target attribute would not point to external words, but for example to time stamps.

3.2 Data Categories

As a second major argument in favour of autonomy, referential markables act as an anchoring point for data categories which are specifically related to anaphoric or (co)referential issues. In addition to the basic identification of a source string,

```
<struct id="m_3"  type="referentialMarkable">
      <struct id="m_1"  type="referentialMarkable">
          <feat type="sourceText" target="w_1..w_2">des technologies</feat>
      </struct>
      <struct id="m_2"  type="referentialMarkable">
          <feat type="sourceText" target="w_7..w_8">une infosphère</feat>
      </struct>
</struct>
```

Figure 5: Representation of a complex markable

it might for example be useful to introduce a data category for the concept of minimal matching string such as proposed in MUC [11], in which a special attribute is used in the answer key to indicate the minimum string that the system under evaluation must include in order to receive full credit for its output.

On the very referential description level, relevant information to be associated with markables pertains usually to characteristics of the referents and is concerned with cardinality, natural gender, definiteness and informational status. /Cardinality/ denotes the size of a set of referents, and /Natural gender/ specifies whether a referent is biologically male or female. It is important to remind that cardinality is distinct from grammatical number, in the sense that it specifies the exact quantity of noun's referent whereas grammatical number is usually vague. In the same way, natural gender is distinct from grammatical gender by the fact that the latter requires agreement between nouns and modifiers, whereas natural gender does not. Since these rules also apply to anaphoric references, annotators frequently observe a discrepancy between gender of an antecedent and a corresponding anaphor. /Definiteness/ whose relations to reference have been extensively discussed in [1] is a category concerned with the grammaticalization of identifiability and non-identifiability of referents on the part of the speaker or the addressee. It takes values such as /Definite identifiability/, /Indefiniteness/, /Generic term/, /Non specific term/ and /Specific term/ and has been used for automatic resolution of definite description by [21]. The relationship between reference resolution and /Informational status/, reflecting the speaker's assumption about the addressee's knowledge, basically expressed as new/old distinction, goes back to [7] and has been applied to anaphora resolution in [15]. More recently, [13] proposed an annotation scheme for information status and annotated a corpus using a set of data categories /Old/, /Mediated/ and /New/, which has been proved to be reliable with respect to inter-annotator agreement. An additional data category, related to the discussion about optional targets, could be /Referential status/, with a conceptual range of /Pending/ or /Solved/ to indicate whether the reference of a referring expression has been calculated or not.

Besides referential information properly speaking, referential markables are often associated with relevant information percolated from lower levels of annotation. Relevant data categories are concerned with morpho-syntactic information such as /Part of speech/, /Grammatical gender/, /Grammatical number/, /Grammatical person/, /Syntactic category/ or /Syntactic function/. Addition-

ally, since reference and anaphora resolution involves heavy knowledge about lexical and semantic properties of the underlying discourse entities, there might be lexical features such as /Abstractness/, /Animacy/, /Collectiveness/, /Alienability/, /Countability/, /Named entity categorization/ or semantic features, such as semantic roles [10].

4 Referential links

4.1 Meta-model component

In addition to markables, any reference annotation schema should be able to express links between source and target markables. Those links represent a relation which is necessary for correct discourse interpretation: depending on the underlying theory, this could be coreference, reference and anaphora. As opposed to a subset of current reference annotation schemes which consider links as a property (or an attribute) of markables [11], we advocate in favour of autonomous links, because this is the simplest way to encode properly ambiguity and multiple antecedents.

From a structural point of view, a /Referential link/ is a RAF meta-model component relating typically one source markable (e.g. the anaphor) to one target markable (e.g. the antecedent). The pointing mechanism is actuated by means of two data categories, /Referential source/ and /Referential target/ (Figure 6). RAF does not allow for more than one target markable per link: reference to a complex antecedent is accounted for by the creation of an embedded markable (see 3.1), whereas ambiguity on target markables is encoded by using an <alt> tag, as proposed in [17] for encoding alternatives (Figure6).

It is currently under discussion whether RAF should consider the target markable to be optional. One possibility for assigning a meaning to links without targets would be the representation of so-called pending reference, i.e. the input to an anaphora resolution system. Another meaning might be the representation of cases where referring expressions *a priori* considered to be involved in links are not related to any previous discourse chunk (first mentions of definite noun phrases, collective pronoun they without antecedent, generic use of nouns, etc.). However, since RAF considers that markables do not necessarily be involved in any link, the question whether those properties are better considered as data categories pertaining directly to the markables remains open. Finally, it has to be noticed that RAF allows one markable to be source and/or target of several links. The first possibility occurs in case of an expression involved in multiple relations, for example a definite noun phrase that is coreferential with a previous expression, but also stands in a bridging relation with another previous expression. The second possibility occurs in case of a same expression being the antecedent for several subsequent mentions.

```
<struct id="m_3"  type="referentialMarkable">
      <struct id="m_1"  type="referentialMarkable">
          <feat type="sourceText" target="w_1..w_2">des technologies</feat>
      </struct>
      <struct id="m_2"  type="referentialMarkable">
          <feat type="sourceText" target="w_7..w_8">une infosphère</feat>
      </struct>
</struct>

<struct id="m_4"  type="referentialMarkable">
      <feat type="sourceText" target="w_13">elles</feat>
</struct>


<struct id="l_1" type="referentialLink">
    <feat type="referentialSource" target="m_4"/>
    <alt>
          <feat type="referentialTarget" target="m_1"/>
          <feat type="referentialTarget" target="m_3"/>
    </alt>
</struct>
```

Figure 6: Ambiguous link (*elles* related to either *des technologies* or to *des technologies ... une infosphère*)

4.2 Data categories

Since the RAF meta-model does not introduce any *a priori* dependency between markables and link, the articulation of links with markables is ensured by the two basic data categories /Referential source/ and /Referential target/. A link may also be assorted with a /Confidence level/ data category. Furthermore, previous work on reference annotation has shown the need of typing the relation between the linked markables. However, as briefly discussed in 1, reference annotation in the sense considered here (covering coreference and anaphora) must face the issue of characterizing properly the types of the relations to be covered. A comparison of relationships involved in current (co)reference and anaphora annotation practice leads to a very heterogeneous inventory, including labels such as identity, coreference, bridging, part-whole, associative, indirect anaphor, unfamiliar, conceptual bridging, set-subset, cause, inferable-of-complement, propositional, possessive, implicit argument, ellipsis, plural NP, numerical pronoun, proper noun, bound anaphor, function-value, instantiation, agent, patient, cause, other-anaphor or role in event [23]. On the other hand, it has been shown for several languages that acceptable inter-annotator agreement could only be achieved on very basic distinctions [14] [22].

Therefore, we propose to start with van Deemter and Kibble's basic distinction of coreference and anaphora [24]. Whereas these two relations may or may not co-occur, current annotation practice generally covers them both by one and the same attribute. MUC-7, for instance, uses the coreference label for any kind of link. On the other hand, the same combination (non-anaphoric coreference, for example) has been characterized differently, depending on whether the framework focus on the relation that holds between the referents of the annotated expressions (coreference) or between the referring expressions themselves (linguistic bridging, NP predication). As a main consequence of these

observations, a consistent annotation framework has to encode two different relationships by two different data categories with different conceptual ranges.

Our principal concern is therefore to introduce a clear distinction between /Objectal relation/ and /Linguistic relation/. Objectal relations are a generalisation of van Deemter and Kibble's extensional approach to the definition of coreference in terms of relations holding between referents: an objectal relation holds between extra-linguistic entities and might be further specified, for example as /Objectal identity/ (in case of coreference), /(In)alienable part of/ or /Subset/. Linguistic relations hold between (parts of) referring expressions (e.g. head nouns) and include for instance /Lexical identity/, /Synonymy/, /Hypernymy/, /Hyponymy/ and /Meronymy/.

As a first advantage, the explicit distinction of objectal and linguistic relations contributes to a more modular task definition, because it enforces to consider separately the question of referential relations and the question of how those relations are realized linguistically. In addition, the simultaneous use of both relationships leads to a more consistent encoding, providing a way of representing simultaneously coreferential and anaphoric relations. It should however be noticed that the precise definition of the conceptual range and the scope of objectal and linguistic relations still needs to be discussed. Open questions are in particular:

- the association of some of the current labels with markables rather than with links (e.g. numerical pronoun, other-anaphor, agent);
- the scope of relationships to be considered (e.g. inclusion of non-anaphoric and non-coreferential relations such as function-value);
- the proper delimitation of objectal and linguistic relations (e.g. instantiation or possession);
- the definition of a set of constraints on combining values for conjoint objectal and linguistic relations (e.g. objectal identity excluding meronymy);
- the granularity of values for conceptual range (e.g. linguistic bridging vs. partitive possessive);
- the inclusion of a specific data category for the direction of an interpretational dependency (anaphor vs. cataphor).

5 RAF and semantic content representation

5.1 RAF as a modelling tool for (co)reference task definitions

RAF framework covers crucial points of structural constraints on reference annotation (autonomous markables and links, representation of embedded markables, ambiguous links, uncertainty). Additionally, it offers a modelling platform for

analysing existing (co)reference task definitions or writing new ones: based on the RAF specifications, the scope of those annotation guidelines it indeed clearly delimited to linguistic questions, i.e. the choice of (a) the type of markables to be covered and (b) appropriate data categories to characterize markables and links, whereas structural issues have no longer to be considered. The combination of these two points guarantees high flexibility with respect to the scope and the semantics of concrete annotation schemes, which are still RAF conform. The first point allows to select any type of source data as markables (raw or pre-annotated texts, spoken language, ...), without constraints on their linguistic realisation (conjoint or disjoint NPs, VPs, ellipses, morphemes, multi-clause units, ...). The second point allows to label markables and relationships with any descriptor available in the data category register.

The concrete decisions as well as their theoretical justification depends on the annotation purpose. However, it has to be noticed that RAF leaves currently open certain points as to what combination(s) of data categories may fit best, for instance, a given project goal. Even if the overall framework allows one to have a background for specifying a scheme and hence facilitates the design of suitable annotation guidelines, one should for example foresee how features may percolate from lower annotation levels, and conversely make sure that what is expressed at the referential level will not contradict further categorization of the corresponding entities at higher level. Typical difficult cases result from the possible combination of morpho-syntactic features, such as /Grammatical gender/, /First person/ and /Grammatical number/, which are strongly correlated with semantic or pragmatic descriptors, respectively /Natural gender/, /Speaker/ and /Cardinality/.

5.2 Beyond reference: another life for RAF

This final section is intended to show how RAF can be used as a basis for the specification of other tasks whose underlying data organization matches the same structural constraints. As an example, one might consider the task of discourse relation annotation, an important pre-requisite for such applications as text summarization, machine translation or reference resolution.

Discourse relation annotation [6] is situated in the general context of the study of the rhetorical structure of texts as developed in Rhetorical Structure Theory [5]. The task consists in identifying texts chunks, usually corresponding to clause level segments, and relating them by discourse relations. These relations are known to be rather ambiguous and relate primary segments as well as higher clusters of segments. Considering this, the framework suggested for reference annotation seems to be appropriate. Indeed, the task presents a strong parallelism with reference annotation. First, it is based on the identification of primary markables which are further organized in a hierarchical way. The structural properties defined for referential markables are indeed consistent with the need of building those complex markables. Second, as (co)reference annotation, it requires the specification of possibly ambiguous links between those markables to instantiate the various discourse relations. It is hence possible to transfer di-

```
<struct type="discourseAnnotation">
      <struct type="discourseMarkable">
            <struct type="discourseMarkable">
                  <feat type="sourceText" id="s1">With its distant orbit - 50
                  percent farther from the sun than Earth - and slim atmospheric
                  blanket,</feat>
                  <feat type="discourseStatus">satellite</feat>
            </struct>
            <struct type="discourseMarkable">
                  <feat type="sourceText" id="s2">Mars experiences frigid weather
            conditions.</feat>
                  <feat type="discourseStatus">nucleus</feat>
            </struct>
      </struct>
      <struct type="discourseLink">
            <feat type="sourceMarkable" target="s2"/>
            <feat type="targetMarkable" target="s1"/>
            <feat type="discourseRelation">background</feat>
      </struct>
</struct>
```

Figure 7: Simple encoding of discourse relations on example 1

rectly the RAF meta-model (Figure 1) to a discourse annotation framework, by re-identifying the relevant components, for instance as *discourse markable* and *discourse link*. Furthermore, the full specification of an annotation format would comprise the following steps:

- select suitable data categories to qualify the discourse markables: one would typically consider categories such as /Segment type/, to characterize the nature of the discourse segment from a linguistic point of view (/Sentence/, /Proposition/, ...), and /Discourse status/ with its two classical values of /Nucleus/ and /Satellite/;
- make a similar selection for links, mainly based on a /Discourse relation/ data category, whose values should be provided by a reference set defined in the data category register, such as /Background/, /Evidence/ or /Explanation/;
- add additional constraints on the combination of the meta-model components and data categories: for instance, as opposed to reference annotation, discourse annotation does not allow for crossed markables and imposes that links only relate markables at the same hierarchical level.

Figure7 shows how this analysis can be instantiated on a simple discourse annotation example (1):

1. With its distant orbit - 50 percent farther from the sun than Earth - and slim atmospheric blanket, Mars experiences frigid weather conditions.

References

[1] Hawkins J. A. *Definiteness and Indefiniteness.* Atlantic Highlands, NJ: Humanities Press, 1978.

[2] Partee B. Opacity, coreference, and pronouns. In D. Davidson and G. Harman, editors, *Semantics of Natural Language*, pages 413–441. Reidel, Dordrecht, 1972.

[3] Partee B. Nominal and temporal anaphora. *Linguistics and Philosophy*, 7:243–286, 1984.

[4] Clouzot C., G. Antoniadis, and A. Tutin. Toward automatic generation of understandable pronouns in french language. *Lectures Notes in Artificial Intelligence*, 1835:242–252, 2000.

[5] Mann W. C. and Thomson S. A. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[6] Marcu D. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 2000.

[7] Prince E. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York, 1981.

[8] Bunt H. and L. Romary. Standardization in multimodal content representation : some methodological issues. In *Proceedings of LREC*, Lisbon, 2004.

[9] Eckert K. and M. Strube. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89, 2001.

[10] Erk K. and S. Pado. A powerful and versatile xml format for representing role-semantic annotation. In *Proceedings of LREC*, Lisbon, 2004.

[11] Hirschman L. and N. Chinchor. Muc-7 coreference task definition. In *MUC-7 Proceedings*, 1997.

[12] Webber B. L. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135, 1991.

[13] Nissim M., S. Dingare, J. Carletta, and M. Steedman. Annotation scheme for information status in dialogue. In *Proceedings of LREC*, Lisbon, 2004.

[14] Poesio M. and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.

[15] Strube M. Never look back: an alternative to centering. In *Proceedings of ACL/Coling*, Montreal, 1998.

[16] Ide N. and L. Romary. A registry of standard data categories for linguistic annotation. In *Proceedings of LREC*, Lisbon, 2004.

[17] Ide N. and L. Romary. International standard for a linguistic annotation framework. *Journal on Natural Language Engineering*, forthcoming.

[18] Modjeska N. Lexical and grammatical role constraints in resolving other-anaphora. In *Proceedings of DAARC*, Lisbon, 2002.

[19] Hobbs J. R. and A. Kehler. A theory of parallelism and the case of vp ellipsis. In *Proceedings of ACL-97/EACL-97*, Madrid, Spain, 1997.

[20] Kibble R. Both sides, now: predictive reference resolution in generation and interpretation. In *Proceedings of IWCS-5*, Tilburg, 2004.

[21] Vieira R. and M. Poesio. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579, 2000.

[22] Vieira R., S. Salmon-Alt, C. Gasperin, E. Schang, and G. Othero. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In Branco A., R. Mitkov, and T. McEnery, editors, *Anaphora Processing*. John Benjamins Publishing Company, 2003.

[23] Davies S. and M. Poesio. Coreference. In *MATE Dialogue Annotation Guidelines, Deliverable D2.1*, 2000.

[24] van Deemter K. and R. Kibble. On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4), 1998.