QurAna: Corpus of the Quran annotated with Pronominal Anaphora

Abdul-Baquee M. Sharaf, Eric S. Atwell
School of Computing
University of Leeds
Leeds, LS2 9JT
United Kingdom
E-mail: a.m.sharaf08@leeds.ac.uk, e.s.atwell@leeds.ac.uk

Abstract

This paper presents QurAna: a large corpus created from the original Quranic text, where personal pronouns are tagged with their antecedence. These antecedents are maintained as an ontological list of concepts, which has proved helpful for information retrieval tasks. QurAna is characterized by: (a) a comparatively large number of pronouns tagged with antecedent information (over 24,500 pronouns), and (b) maintenance of an ontological concept list out of these antecedents. We have shown useful applications of this corpus. This corpus is the first of its kind covering Classical Arabic text, and could be used for interesting applications for Modern Standard Arabic as well. This corpus will enable researchers to obtain empirical patterns and rules to build new anaphora resolution approaches. Also, this corpus can be used to train, optimize and evaluate existing approaches.

Keywords: Anaphora, Quran, Corpus

1. Introduction

In this paper, we report the release of QurAna: a corpus of the Quran annotated with antecedent references of pronouns. We describe in detail the annotation scheme, annotation process, and intended use of this language resource. We believe this annotation work on Classical Arabic is the first of its kind and will be a valuable language resource for the corpus linguistics community. This corpus will benefit researchers in extracting empirical patterns and rules for building new anaphora resolution approaches. Also, this corpus can be used to train, optimize and evaluate existing approaches.

The ability to identify the antecedents of a pronominal anaphor has many applications in NLP. For example, correct antecedence assignment is vital for efficient machine translation. Similarly, information extraction and question answering systems can work efficiently when pronouns are resolved correctly. Also, automatic summarization systems rely on producing cohesive meaning through proper anaphor chaining. The importance of this task was recognized through the Message Understanding Conference (MUC) community (Hirschman & Chinchor, 1997) with a separate task for developing automatic coreference resolution systems.

Despite increasing work on corpus building in recent years and recognition of the importance of annotated corpus data for various tasks, annotation of anaphoric links is still lagging behind (Mitkov et al. 2000).

1.1 Why this text?

The Quran is a scripture which is according to Muslims the verbatim words of Allah containing over 77,000 words revealed through Archangel Gabriel to Prophet Muhammad over 23 years beginning in 610 CE. It is divided into 114 chapters of varying sizes, where each chapter is divided into verses, adding up to a total of 6,243 verses.

The Quran was chosen for creation of this annotated corpus for a number of reasons:

1. The original Arabic Quran is characterized by very frequent use of anaphors. The majority of anaphoric devices in the Quran appear around pronominal anaphora. Hence, the ability to resolve pronoun antecedence is vital to understanding the Quran. We will demonstrate pronominal anaphora in the Quran further through examples from Quranic verses in the next sub-section.

2. The Quranic scripture is a widely used and cited document that guides the lives of over 1.5 Muslim adherents today. Increasingly non-Arabic speaking Muslims – and many non-Muslims – learn Classical Arabic with the objective to understand the Quran. For Arabic speakers, the Quran is considered to be the finest piece of literature in the Arabic Language. Producing language evaluation resources for computational analysis of pronominal anaphora of such an important text should be well justified.

3. Classical Arabic (CA) text is the form of Arabic language used in literary texts authored by early Arabic scholars mainly in the 6th through 10th century. The Quran is considered to be the highest form of Classical Arabic text and has been extensively cited in linguistic scholarly works since the 7th century. In contrast to most languages, the total body of Arabic texts published during this classical period large, compared to modern corpora of Modern Standard Arabic (MSA) - the form used in contemporary scholarly published works as well as in the media. MSA does not differ from Classical Arabic in morphology or syntax, but richness of stylistic and lexis usage is apparent in Classical works. This makes Classical Arabic subsume MSA making computational and linguistic research work on CA benefit both. However, most recent work on Arabic corpus annotation has concentrated on MSA, and the computational corpus linguistic community has largely ignored study of the large body of language resources available in Classical Arabic.
4. Being a central text in Arabic, over the past 14 centuries a large body of scholarly commentary volumes has been compiled elaborating on linguistic, stylistic, semantic and other aspects of the Quran. This makes the task of compiling evaluation datasets and annotated corpora on the Quran simpler; as it is very likely we can find scholarly comments on any difficult annotation question. In our QurAna corpus, we relied on assigning correct antecedence of pronouns in ambiguous cases on scholarly commentary of Ibn Kathir - a well-known Quranic scholar who died in 1373 CE.

5. The Quran is widely translated into almost all live languages of the world, and in many cases multiple translations within one language are available. Among these translations a number of them are also available in machine readable electronic format in the web. All translations maintain chapter and verse numbers as available in the original text, allowing alignment between these translations at sentence or verse level. Moreover, as the Quran is believed to be the words of God, all translations are made very carefully. Given this fact, any language evaluation resource in the source language of the Quran could be used to evaluate computational tasks on the target language translation as well. Having the source language properly annotated with pronoun antecedents enables evaluation of other language translations as well as evaluation of rival translations in one language. Moreover, there has been analogous research to produce a corpus of the Bible aligned in many translations (Resnik et. al, 1999).

6. The size of the Quran is manageable for manual or semi-automatic annotation tasks. Given that the Arabic language still lacks many NLP resources available for a language like English (e.g., tagsgers, parsers, Wordnet, frameNet, etc.), developing manually annotated language resource on a smaller scale like the Quran text could be a good starting point. A case in point is the Quranic Arabic Corpus project (Dukes, et al 2010), where every word of the Quran is tagged with morphological, part-of-speech and syntactic information, and is publicly available for research purposes. Another available resource is QurSim, a corpus of the Quran annotated with related verses (Sharaf & Atwell, 2012). Our QurAna corpus, along with these other available resources on the Quran, will enable interesting computational linguistic applications on the Quran which in turn will eventually create motivation for wider applications and resource development for Classical and Modern Standard Arabic.

7. With the wide spread of knowledge in machine readable formats (e.g., ontologies, wikis, corpora, digitized libraries, etc.), and the availability of large bodies of Arabic texts from both the classical as well as modern period, we find increasing interest in incorporating world knowledge in information and knowledge extraction tasks (Gabrilovich & Markovitch, 2007). As the Quran frequently uses pronouns that require such domain knowledge to resolve their antecedents, we think this resource would be very valuable for researchers in this new direction.

1.2 Pronominal Anaphora in the Quran

The Quran is characterized by frequent and varied use of pronouns, at least in part because the morphology of Arabic in general requires a pronoun with each verb, which is not the case in many languages including English. This is evident in the following example from verse (24:31). Note that the English translation retains only 16 pronouns from the original Arabic’s 27 pronouns. Also, note that the underlined Arabic pronouns are affixed to the word rather than standing alone.

And say / to the believing women / (that) they should lower / [of] / their gaze / and they should guard / [their] chastity, and not / they (to) display / [their] adornment / except / what / is apparent / of it. / And let them / their head covers / over / [their] bosoms, / and not /they (to) display / [their] adornment / except / to [their] husbands, / or / [their] fathers / or / fathers / (of) / their husbands / or / [their] sons / or / sons / (of) / [their] husbands / or / [their] brothers / or / sons / (of) / [their] brothers / or / sons / (of) / [their] sisters, / or / [their] women / or / what / possess / [their] right hands / or / the attendants / having no physical desire / among / [the] men / or / [the] children / who /they (are) not / aware / of / private aspects / (of) / the women. / And not / let them / their feet / to make known / what / they conceal / of / [their] adornment.

Also, the following verse shows intensive reliance on pronouns.

And indeed, they / surely, turn them / from / the Path / and they think / that they / (are) guided

And lo! they surely turn them from the way of Allah, and yet they deem that they are rightly guided [43:37]

1 See for example: http://quran.com and http://qurandatabase.org
2 http://corpus.quran.com
3 http://textminingthequran.com/wiki
4 We follow the Arabic text with English word-by-word translation available at http://corpus.quran.com, followed by Pickthall translation available at http://quran.com - a subset of pronouns are underlined to illustrate the point in the text
Resolving pronoun reference is vital to understanding the meaning of the Quran. Consider the following two verses, the underlined pronouns refer to “children of Israel” which has been tagged as such, although the actual mention of the antecedent was made very earlier in the text.

And when / We took / your covenant / and We raised / over you / the mount.
And when We made a covenant with you and caused the mount to tower above you [2:63]

And when / We raised / the mountain / above them
And when We shook the Mount above them [7:171]

The Quran relies on the reader’s world knowledge and intuition when using pronouns without explicitly including any antecedent information. For example a good number of second person pronouns in the Quran refer to Prophet Muhammad with no prior mention of his name, as in verse 2:4 below.

In terms of counting pronouns in the Quran, we relied on the Quranic Arabic Corpus (QAC) to produce the statistics presented in table 1 below.

|     | Total count | %    |
|-----|-------------|------|
| Person |          |      |
| 1st person | 3,903 | 13.3%  |
| 2nd person | 6,881 | 23.3%  |
| 3rd person | 13,933 | 47.2%  |
| None/other | 4,777 | 16.2%  |
| Gender |          |      |
| Masculine | 22,284 | 75.6%  |
| Feminine | 1,822  | 6.2%   |
| None/other | 5,388 | 18.3%  |
| Number |          |      |
| Singular | 9,141 | 31%    |
| Dual | 381  | 1%     |
| Plural | 17,671 | 60%   |
| None/other | 2,301 | 8%    |

Table 1. Distribution of pronouns in the Quran

2. Quranic Pronoun Reference Annotation

2.1 Related Annotation Schema

The first anaphora annotation scheme is the Lancaster IBM project at UCREL (Garside et al. 1997). Under this scheme the antecedent (whether anaphor or cataphor) is enclosed in brackets and given an index number and the proform (i.e., the pronoun) is preceded by the ‘REF’ symbol with the index number along with either ‘<’ or ‘>’ symbol indicating the direction: either anaphora or cataphora. For example:

(6 the married couple 6) said that <REF=6>they</REF> were happy with <REF=6>their</REF> lot.

This scheme was used to annotate part of the AP corpus consisting around 100,000 words.

Another tagging schema is MUC-7 SGML schema (Hirschman and Chinchor 1997) which accompanied the task definition of the MUC-7 on coreference annotation. The following is a sample annotation from this corpus:

The GNOME project relies on an earlier general purpose annotation scheme called MATE. This scheme was designed keeping in mind a ‘discourse model’ and thus aimed at annotating ‘discourse entities’ and any co-reference to them. Under this scheme <de> is the main discourse element, and <link> is used to mark information about anaphoric relations using <anchor> elements. Here is an example (Poesio 2004).

AQA (Boldrini et al. 2009) is a multilingual anaphora annotation scheme that can be applied in machine learning for the improvement of Question Answering systems. This scheme has been used to annotate the CLEF 2008 corpus in Spanish. There are several markups used to specify anaphora type (e.g., pronominal, superficial, adverbiaal, ellipse and definite descriptions) and others to specify the relation type between anaphoric expression and its direct or bridging antecedent. The following is an example, where <t> = topic, <subt> = subtopic, <q> = question, <de> = discourse entity, <link> = anaphora, <rel> = relationship, <status> = sure or uncertain, <ant> = antecedent, <refq> = question-answer pair:
| QurAna (our corpus)                          | Classical Arabic | Corpus Size | # of pronouns |
|--------------------------------------------|-----------------|-------------|---------------|
| (Weischedel and Brunstein 2005)            | English/Penn    | 1 million   | 24,104        |
| (Ge, Hale and Charmani 1998)               | English         | 93,931      | 2,477         |
| (Hasler et al 2006)                        | NP4E/English    | 55,000      | 2,100         |
| (Barbu 2003)                               | English         | 55,000      | 653           |
| (Hervas and Finlayson 2010)                | English         | 24,422      | 7,207         |
| (Barbu 2003)                               | French          | 36,000      | 482           |
| (Hammami and Belguith 2011)                | Arabic          | 164,051     | 4,300         |

Table 2. Comparison of QurAna with other available corpora for anaphora resolution

Using this scheme a pilot evaluation corpus was manually annotated out of the CLEF multilingual corpus with 600 questions: 200 for each English, Italian and Spanish with an average agreement of 87%.

(Hammami et al 2009) presents a corpus annotated with coreference chains for Arabic using a custom-designed XML tool called AnATAr. This corpus is of size 77,457 words (very close to the size of the Qur’an) and includes newspaper articles, technical manual, a book on education and a novel. The scheme is adopted from (Tutin et al 2000) and is compatible with MUC scheme. Here is an example output.

2.3 Annotation Process
Guided by the previous annotation schema, and by the nature of Arabic usage of pronouns, and in particular domain specific usage of pronouns in the Quran, we started the annotation process. First, the 128,000 word segments of the Quran were maintained in a MySQL database and unique IDs were assigned to each word segment. Next, QAC was used to identify those targeted segments that contain pronouns, and for each pronoun the starting and ending IDs of the text span that represents antecedents were recorded manually through forms developed using PHP scripting language. We chose QAC as it underwent various verification levels for accuracy of part-of-speech annotation, and is placed for online collaboration for further validation. We took the instances of pronouns that were tagged as ‘PRON’ in the QAC corpus. This tag covers all kind of personal pronouns: 1st, 2nd and 3rd persons; singular, dual and plural; connected pronouns (where the personal pronoun is suffixed with noun or verb) and separate stand-alone pronouns. However, we left out demonstrative and relative pronouns as they are less in number (approx. 15%) and often their antecedents are non-anaphoric.

As we progressed with tagging pronoun antecedents in the Quran, we maintained a concept ontology out of these antecedents. In total we gathered over 1050 concepts. For example, a concept in our list is “Quran” and all pronouns referring to “Quran” were linked to this concept regardless of the presence or absence of the actual antecedent in the immediately surrounding text, or even various different names by which the concept “Quran” is referred to in the Quran: like Kitab (book), Dhirk (remembrance), al-Furqan (the criteria), and such other attributes. Similarly, all pronouns referring to prophet Muhammad were linked although the actual antecedents might have different words like Rasoul (messenger) or Nabi’ (prophet). Keeping an ontology of concepts is very helpful for information retrieval and many other useful applications. Table 3 (next page) gives the 20 most frequently referred-to concepts in our list.
/ the world / and the Hereafter, / then let him extend / a rope / to / the sky, / then / let him cut off / then let him see / whether / will remove / his plan / what / enrages. / Whoso is wont to think (through envy) that Allah will not give him (Muhammad) victory in the world and the Hereafter (and is enraged at the thought of his victory), let him stretch a rope up to the roof (of his dwelling), and let him hang himself. Then let him see whether his strategy dispelleth that whereat he rageth!.

This manual annotation was done by the first author and it took him over one year to annotate a total of 24,679 pronouns that cover the entire Quran.

2.4 QurAna corpus
Following the process described above, the entire Quran was annotated. Table 4 below gives a quantitative account of key statistics of this corpus. Note that the 24,679 tagged pronouns include anaphoric as well as non-anaphoric cases, and relative and demonstrative pronouns are excluded.

### Table 4. Quantitative measures from QurAna corpus

| Measure | Count |
|---------|-------|
| # of word segments | 127,795 |
| # of pronouns | 24,679 |
| # of third person pronouns | 11,544 |
| # of sentences | 6,236 |
| Average Distance between pronoun and antecedent | 30 word segments |
| % of antecedents within the same sentence as the pronoun | 56% |
| % availability of antecedents | 54% |
| Total number of concepts | 10,545 |

The majority of our pronouns are anaphors, and only 99 instances (0.3%) showed cataphoric relation where the antecedents were mentioned after the pronoun. Although in certain cases the antecedent is mentioned way back, in the majority of cases they are found within 200 word segments from the pronoun. Among the total 13,158 pronouns which have antecedents, only 2,309 (17.5%) antecedents matched with the nearest preceding noun. Considering the whole population of pronouns only 9% of antecedents are captured correctly when attached with the nearest preceding noun. Among the total 2nd person singular pronouns 27% of them referred to Prophet Muhammad.

This corpus is made public for both online query – as described in the following section - or for download.

3. Applications using QurAna

#### 3.1 Online Visualization

Using PHP scripting and access to the annotated corpus captured as a MySQL database, a number of query pages

---

5. [http://www.textminingthequran.com/wiki/Pronoun_Reference_in_the_Quran](http://www.textminingthequran.com/wiki/Pronoun_Reference_in_the_Quran)
were made online⁶. Entering a verse number, a user can get all pronouns along with their antecedence and all concepts this verse has. Figure 1 below gives an example screenshot from the online query page. The actual verses are quoted in Arabic, however, the verse number leads to English – and potentially many other language translations through hyperlinks to an external site.

Figure 1. pronoun resolution of verse 38:29

The user may explore from the concepts listed for this verse, to all other verses that share this same concept, represented as concordance lines for convenient analysis. Figure 2 shows instances where the concept ‘mankind’ is repeated in the Quran as pronouns.

Figure 2. concordance lines for the concept ‘mankind’

### 3.2 Verse Distance using Vector Space Model

The vector space model is widely used in information retrieval where the query terms and each document are represented as vectors and the distance between query and document is measured by comparing the cosine of angle between them. We followed the same methodology and considered each verse of the Quran as a separate document.

Each verse then was modeled as a term vector taking roots of the Quran as the terms. The Quran has 1,226 unique roots, from these we have kept roots repeated over 2 times, and removed the first 3 most frequent roots. Thus, our vector for each verse contains 758 roots as term indices. Next, in order to give a weight for each term, we used term frequency – inverse document frequency (tf-idf) metric, using the following formula adapted from (Sebastiani 2002):

\[
tfidf(t_k, d_j) = \frac{\#(t_k, d_j)}{\log \left[ \frac{|T_r|}{\# T_r(t_k)} \right]}
\]

Where \(\#(t_k, d_j)\) denotes the number of times the root \(t_k\) occurs in the verse \(d_j\), and \(\# T_r(t_k)\) denotes the verse frequency of root \(t_k\), that is, the number of verses in the Quran \(T_r\) in which the root \(t_k\) occurs.

In order for the weights to fall in \([0,1]\) interval and for the verses to be represented by vectors of equal length, the weights \((w_{kj})\) resulting from tfidf were normalized according to the following formula for cosine normalization:

\[
w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{i=1}^{n} (tfidf(t_k, d_i))^2}}
\]

To find the distance (or measure of similarity) between two vectors, cosine angle is measured using the formula below, where \(A, B\) denotes two verses’ vectors:

\[
similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}
\]

Similarity values fall between \([0,1]\), where 0 indicates no similarity, and 1 indicates identical matching. Using the above setup, we have evaluated a dataset of 7,679 related verse pairs we created from scholarly works (Sharaf and Atwell 2012) and found out that only 428 pairs (6%) produced similarity value above 0.5. This finding confirms the assumption that automatic computation of verse relatedness requires integration with domain specific knowledge source and relying only on lexical matching produces poor results.

Given these results we considered next how to enrich a verse vector with concepts from our ontology. Instead of constructing root vectors for a verse from only that verse’s root, we augmented this verse’s roots with roots of all other verses that share common antecedent. For example consider verse 27:26 below:

\[
\text{Nay, / is arrested / their knowledge / of / the Hereafter? / Nay / they / (are) in / doubt / about it. / Nay, / they / about it / (are) blind. / Nay, but doth their knowledge reach to the Hereafter? Nay, for they are in doubt concerning it. Nay, for they cannot see it.}
\]

http://www.textminingthequran.com/apps/pron.php

---

⁶ http://www.textminingthequran.com/apps/pron.php
This verse contains 3 concepts marked by pronoun referents: ‘the polytheists’, ‘those who deny resurrection’ and ‘the world Hereafter’. Therefore, we have augmented the term vector of the verse 27:66 with the terms from all other verses that have any of these three concepts.

The similarity measurement experiment described above was repeated using these improved vectors, and the same dataset was used. While in the early experiment, only 428 pairs showed similarity distance over 0.5, augmenting verses with their concepts showed 869 pairs from the total of 7,679 pairs in our dataset, i.e., over 50% improvement.

4. Challenges

We have encountered a number of challenges while pursuing this task. Often, the distance between the pronoun and its antecedent is too far. This is evident more in the case of long stories, where the main characters might be mentioned only once at the beginning and all subsequent references are done through pronouns. For example, in Chapter 2, a series of verses addressed the ‘Children of Israel’ where explicit mention is made at the beginning of the dialogue but most subsequent reference are made through 2nd person pronouns sometimes as far as 33 verses away. Also, as our annotation scheme does not allow discontinuous antecedents or multiple antecedents, in such cases we had to include as antecedents the whole text span, resulting in some compound concepts.

Often the Quran makes grammatical shifts deliberately for various purposes (for example to draw attention), and as a result the number or person agreement between the pronoun and the antecedent is violated. Consider for example verse 65:1 where the singular noun antecedent ‘prophet’ disagrees with the plural 2nd person pronoun used (you):

والَأَيُّهَا الَّذِيْنَ أَنْعَمَ اللَّهُ عَلَيْهِمْ مِنْ نَجَاتٍ وَخَصْصَوْا الْعَلَّةُ

O Prophet! / When / you divorce / [the] women, / then divorce them / for their waiting period.

“O Prophet, when you [Muslims] divorce women, divorce them for [the commencement of] their waiting period and keep count of the waiting period…”

There were a number of challenges faced while tagging pronouns with a concept name. Often a decision to create a new specific concept or maintain an already available generic concept was required. For example, in verse 67:5 the word ‘lamp’ was used to mean ‘stars’, and hence pronouns could be tagged with either of these two concepts. In this particular case, we decided to tie the pronoun to the concept ‘star’ rather than the concept “lamp” so that all verses referring to ‘star’ can be linked plus as we keep reference to actual antecedence, we still can retrieve that stars are referred to in the Quran as lamps.

5. Conclusion

We have presented QurAna as a language resource for Quranic scholars, students and for researchers in the computational linguistics community, particularly those investigating computational anaphora resolution systems. We have tagged over 24,000 Quranic pronouns with their antecedence information. QurAna is characterized by: (a) comparatively large number of pronouns tagged with antecedence information, and (b) maintenance of an ontological concept list out of these antecedents. We have shown useful applications of this corpus. This corpus is the first of its kind considering Classical Arabic text, and would find interesting applications for Modern Standard Arabic as well.

6. References

Abbas, N., Atwell, E. (2012). Qurany: how to search for concepts rather than words in a corpus. Proc IVACS’2012, Leeds, UK.
Barbu, C. (2003). Bilingual Pronoun Resolution: Experiments in English and French. PhD Thesis. Univeristy of Wolverhampton, 2003.
Boldrini, E., Puchol-Blasco, M. Navarro, B., Martínez-Barco, P., Vargas-Sierra, C. (2009). AQA: a multilingual Anaphora annotation scheme for Question Answering. Procesamiento del Lenguaje Natural, Revista nº 42
Dukes, K.; Atwell, E.; and Sharaf, A. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. Proc LREC’2010, Valetta, Malta.
Gabrilovich, E.; and Markovitch. S., (2007). Computing semantic related-ness using wikipedia-based explicit semantic analysis. In Proceedings of the 20th International Joint Conference on Arti’cial Intelligence, January.
Garside, R., Fligelstone, S. and Botley, S. (1997). Discourse annotation: anaphoric relations in corpora. In Corpus Annotation, Pearson.
Ge, N., Hale,J., and Charniak, E. (1998). A statistical approach to anaphora resolution. In Proceedings of the Sixth Workshop on Very Large Corpora, pages 161–170.
Hammani, S. and Lamia Hadrich Belguith. النحل (2011). التحليل للأصنام اللغوية والألفاظية، جامعة الامام محمد بن سعود للعلوم الإنسانية، الرياض، المملكة العربية السعودية, International Computing Conference in Arabic (ICCA’2011), Riyadh-Saudi Arabia, 31 may-2 june 2011
Hammani, S., Belguith, L. And Hamadou, A. (2009) “Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links”. The International Arab Journal of Information Technology, Vol. 6 No. 5, pp.
Hasler, L., Orasan, C., and Naumann, K. (2006) NPs for Events: Experiments in Coreference Annotation. In Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006), 24 -- 26 May, Genoa, Italy, pp. 1167 -- 117

Hervas, R. and Finlayson, M. (2010). “The prevalence of descriptive referring expressions in news and narrative.” In Proceedings of the ACL2010 Conference Short Papers, Uppsala, Sweden, July 2010.

Hirschman, L. and Chinchor, N. (1997). MUC-7 coreference task definition. In MUC-7 Proceedings. Science Applications International Corporation.

Poesio, M. (2004). "The MATE/GNOME Scheme for Anaphoric Annotation, Revisited", Proc. of SIGDIAL, Boston.

Resnik, P.; Olsen, M.B.; and Diab, M. (1999) The Bible as a Parallel Corpus: Annotating the `Book of 2000 Tongues, Computers and the Humanities, 33(1-2), pp. 129-153.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Comp. Survey, 34(1):1-47.

Sharaf, A., Atwell, E. (2012) QurSim: A corpus for evaluation of relatedness in short texts. LREC 2012, Istanbul, Turkey.

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., and Antoniadis G. (2000) “Annotating a Large Corpus with Anaphoric Links,” in Proceedings of the Discourse Anaphora and Reference Resolution Conference. pp. 134-137, UK, 2000.

Weischedel, R. and Brunstein, A. (2005) BBN pronoun coreference and entity type corpus. Linguistica Data Consortium.