Interactive Corpus Annotation of Anaphor Using NLP Algorithms

Catherine Smith[1] and Matthew Brook O'Donnell[1]

1. Introduction

Pronouns occur with a relatively high frequency in all forms English discourse. The nature of Greek as an inflected language means that pronouns and zero anaphora in verbs occur with an even greater frequency. Resolving this anaphora is fundamental to the understanding of any language but is generally unresolved in corpus data and therefore a large amount of potentially useful data for corpus methods such as concordance and collocational analysis is lost. In order to be able to access the data 'hidden' in anaphora the corpus needs to be annotated with the anaphoric relations. This is a time consuming task which would ideally be done automatically or at least interactively with the computer presenting candidate values for the annotator to select from. However, although identifying the antecedent of a pronoun in text is usually an easy and unconscious task for a human interpreter, it has proved to be one of the more challenging tasks for natural language processing systems. This paper focuses on particular on the development of a computer aided participant annotation system based on computational techniques. It has been developed for use with the OpenText.org corpus of Hellenistic Greek but the principles are relevant to any language.

1.1. The OpenText.org corpus

The aim of the OpenText.org project is to build a linguistically annotated corpus of Hellenistic Greek[2] to aid the study of the New Testament. For practical reasons the project has so far focussed primarily on the New Testament (around 130,000 words). The corpus has been manually annotated at several levels using a framework adapted from Systemic-Functional Linguistics including grammatical information at the word level and clause level structures using Subject, Predicator, Complement and Adjunct slots. An overview of the annotation model used can be seen in figure 1. The project is currently focussing on the annotation of participants in the corpus which play a large role in the interpersonal metafunction, the particular points in the annotation model are shown in bold in figure 1.

The files underlying the OpenText.org corpus are xml files. To aid a staged approach to annotation (one linguistic level at a time) and to make the files easier to maintain and less complex a standoff markup system is used. This means that each level of annotation only stores the information for its own level or in some cases only for part of a level. In addition each text in the corpus has one xml file which combines some central elements of the annotation available for that text. These combined files

[1] University of Liverpool
e-mail: m.odonnell@liv.ac.uk

[2] Hellenistic Greek can be defined as the Greek used in the Hellenistic and Roman worlds from around the fourth century BCE to fourth century CE (O'Donnell 2005, 3).

are constructed from the separate xml files and are used for searching and for constructing web interfaces as combining the separate xml files each time is to slow to be practical for those tasks. An example of the combined xml file is shown in figure 2. The use of separate annotation files means that the output of the participant analysis tool does not need to be merged with any other xml files but can rather be independent using href attributes to provide the link to the word id's in the base files. An example of the participant output is shown in figure 6, section 3.

The text chosen as the example text for this paper is 3 John. This is an epistolary text of 219 words. It is one of the smallest texts in our corpus and is one of two passages selected as development material for the algorithm (the other being a section of a narrative text from the Gospel of Mark). A literal translation of 3 John with the discourse referents underlined can be found in appendix 1.

	Field	Tenor	Mode
Pericope	Semantic Domains; Process patterns; Circumstance; patterns; Aspect patterns; Causality patterns	Participants and reference types; Attitude patterns; Person reference patterns	Clause level boundaries; Theme
Clause	Structural Summary: SFPCA		
	Process and Participants; Aspect; Causality	**Participants Involved**; Attitude	Theme and Rheme; Clause Boundaries
Word Group	Structure: head term, specifier, definer, qualifier, relator		
	Semantic Domain of Head Term	**Type of Participant Reference** (grammaticalised, reduced, implied)	Word Group Boundaries

Figure 1: A summary of language features by rank and metafunction (Smith, 2005: 136; adapted from O'Donnell, 2005: 169-70).

```
<cl.clause xml:id="NT.3Joh.1_c54" level="primary" connect="NT.3Joh.1_c53"
                                                                structure="S-C">
       <cl.S>
           <wg.group xml:id="NT.3Joh.1_wg151">
               <wg.head>
                   <wg.word xml:id="NT.3Joh.w209" ref="NT.3Joh.1.15">
                       <pos>
                           <NON num="sing" cas="nom" gen="fem"/>
                       </pos>
                       <wf betaLex="ei)rh/nh" betaForm="ei)rh/nh"
                                                        lex="ε□ρ□νη">ε□ρ□νη</wf>
                       <sem>
                           <domain majorNum="22" subNum="42" select="1"/>
                           <domain majorNum="25" subNum="248"/>
                       </sem>
                   </wg.word>
               </wg.head>
           </wg.group>
       </cl.S>
       <cl.C>
           <wg.group xml:id="NT.3Joh.1_wg152">
               <wg.head>
                   <wg.word xml:id="NT.3Joh.w210" ref="NT.3Joh.1.15">
                       <pos>
                           <PRO num="sing" cas="dat" per="2nd" type="per"/>
                       </pos>
                       <wf betaLex="su/" betaForm="soi" lex="σ□">σοι</wf>
                       <sem>
                           <domain majorNum="92" subNum="8"/>
                       </sem>
                   </wg.word>
               </wg.head>
           </wg.group>
       </cl.C>
   </cl.clause>
```

Figure 2: An example of an Entry from the Combined File.

2. Computational approaches to the problem

The problem of pronoun resolution has been a significant research area for computational linguistics since the 1970s. Several approaches to the task have been considered and implemented. These approaches are split into two broad categories, those that rely on statistical evidence also known as knowledge-poor, and those that use some form of discourse model of text, knowledge-rich (Mitkov, 1999; Deoskar, 2004). Each approach has its advantages and disadvantages. Algorithms falling into the first category require large corpora of data but can work with sparsely or even un-annotated text. They tend to use parsing tools and a training corpus together with machine learning or genetic algorithms to build up a statistical picture of the language usage which provides the background on which to make decisions regarding anaphora. These approaches then, do not require the user to provide information about language patterns or discourse structures 'up-front' but rather use learned probabilities to perform their task. In contrast, knowledge-rich approaches explicitly encode several linguistic phenomena relating to anaphoric references. This information is generally supplied in the form of rules and can include syntactic and/or discourse based information. This requires the user to provide detailed information about the language and about anaphora patterns before it can be of any use. This approach can also often be more reliant on the accuracy of the grammatical parsers used to pre-process the text.

When implemented for English both approaches record accuracy rates of up to eighty-eight percent although these high levels of accuracy are limited to very specific

genres of text for which the algorithm has either been specially written for or been specially trained on (Hobbs, 1978: 342-5; Walker, 1989: 254; Lappin and Leass, 1994: 554, 556; Tetreault, 1999: 604; Okumura and Tamura, 1996: 875). For more general language application the results are closer to the fifty to sixty percent mark (Mitkov *et al.*, 2007). This is again for both approaches although there are more knowledge-rich approaches achieving figures in that region that there are knowledge-poor approaches.

For application to our small but already richly annotated corpus of Ancient Greek a knowledge-rich approach is the more appropriate. There is not enough text to provide sufficient data for statistical approaches but it does contain accurate hand annotated details of linguistic information for individual words, word group structures and clause structures, which are of huge relevance to the knowledge-rich approaches.

As knowledge-rich approaches to anaphora have developed, different features of the language have provided the 'knowledge' for the algorithms. One of the earliest approaches, which is still well regarded today, was that of Hobbs (1977; 1978). The algorithm uses a simple breadth-first search of the syntax tree which stops once a noun phrase which grammatically agrees with the pronoun is found. If no potential antecedent is found in the current sentence the algorithm moves to the previous or parent sentence and repeats the same technique.

The order in which the tree is searched favours certain antecedents and is the key to the algorithm. Because the immediate NP or S is searched first followed by each previously occurring one, recent referents are favoured over those further back in the text. The left-right breadth first search also favours certain grammatical roles. Subject roles are favoured over object roles because of SVO English word order and the left-right bias, whereas the Breadth first search favours objects over adjuncts because noun phrases in prepositional phrases are more deeply embedded in the phrase structure than are objects. This reliance on order means that the algorithm can only be usefully applied to SVO order languages, which Greek is not.

Other approaches use a discourse model rather than a syntax tree as their main source of information. One such approach is centering theory which uses the basic premise that only one discourse entity is in focus at once (Brennan *et al*., 1987). The algorithm relies on the related ideas that the entity which is in focus or 'centered' is more likely to remain the focus of future utterances and that this entity is more likely to be pronominalised than any other (Deoskar, 2004: 5). Another approach is Salience theory which uses both syntactic structures and the concept of attentional state (similar to a 'center') but has no explicit discourse model (Lappin and Leass, 1994: 535). The algorithm works on the output of McCord's Slot Grammar (McCord, 1980). It gives different weights to a variety of grammatical features in potential candidates and the one with the highest weight is taken as the antecedent.

Although Centering theory has been tested on a variety of languages including modern Greek, Salience theory is the more logical choice for Greek. The algorithm is more transparent and therefore more easily optimised for the corpus. In addition the slot grammar framework that underlies Salience theory has much in common with systemic grammars (McCord, 1980: 31) which form the basis of the OpenText.org annotation on which the algorithm will be required to work.

2.1 Salience theory

The resolution algorithm in Salience theory is quite simple but in addition to this resolution algorithm a series of filters are also required which handle tests for morphological agreement, co-reference and pleonastic pronouns and which are used to identify potential candidates (Lapin and McCord, 1990a; 1990b; Lappin and Leass, 1994: 536). The resolution algorithm works as follows. For each discourse entity in a sentence a salience weight is calculated based on the weightings given in figure 3. The entities are added to the salience model one at a time in text order. At the end of each sentence all weightings are halved and the scores from the next sentence are added to the new total. This ensures that recency is prioritised but does not restrict focus to one entity. When a pronoun is encountered all possible antecedents (based on the filtered data) are selected from the full list. At this point two more phenomena aretaken into account. If the proposed antecedent performs the same grammatical role as the pronoun 35 is added to its weight (role parallelism). If choosing the entity

Sentence Recency	100
Subject Emphasis	80
Existential Emphasis	70
Accusative (Direct Object) Emphasis	50
Indirect Object and Oblique Complement Emphasis	40
Non-Adverbial Emphasis	50
Head Noun Emphasis	80

Figure 3: Salience Factors in Lappin and Leass's System (Jurafsky and Martin, 2000: 685).

results in the pronoun being a cataphoric rather than anaphoric reference then -175 is added to the score (heavily favouring anaphora). Once these scores have been added the weightings are compared and the entity with the highest weighting is selected as the antecedent. When an antecedent is identified its salience weights are added to the totals but the parallelism and cataphor weights are ignored. If an entity occurs twice in the same sentence only its highest score is counted.

2.2 Implementation for Ancient Greek

While the basic algorithm described in section 2.1 has been retained in this Ancient Greek implementation there are some necessary changes and adaptations which are described here. In addition the grammatical and syntactic filter element of the algorithm requires a different implementation. Ancient Greek has a high level of inflection so the filtering system is able to play a larger part in the algorithm as the three-gender system reduces the number of candidates from which the resolution algorithm must select. It does, however, also means that anaphora is carried not just by pronouns but also by verbs.

2.1.1 Identifying discourse referents

Salience theory requires all discourse referents to be given a salience weight in order to build up a picture of the shifting focus, therefore it is necessary first to identify these discourse referents. Although this is reasonably straightforward when reading through a text it is not an easy task to accomplish algorithmically. At present the algorithm identifies words as discourse referents which exhibit one of the following characteristics:

- Nouns
- Adjectives (in Subject or Complement slots, with article or in Vocative case)
- Participles (with article)
- Finite verbs (due to the person and number inflection)
- Pronouns (discounting interrogatives)

2.1.2 Grammatical agreement

The first and probably most important of the filters needed for reference resolution is grammatical agreement. In Greek this involves testing against three different systems, gender, person and number. The OpenText.org corpus already contains annotation for these systems so checking grammatical agreement is reasonably straightforward. Each chain of referents keeps a record of person case and number or records it as not known if none of the instances exhibit the feature. As new instances of the participant chain are added (by either resolution or, in the case of nouns and adjectives, string matching) any missing values are added if they are present in the new occurrence. In order to be considered an agreement words must match with the referent chain in any systems which are recorded or be missing the value itself, so for example a third person masculine plural verb could match a chain having the values of third person, masculine and plural or one with third person and masculine but without any assigned value for number. If the verb was subsequently resolved to that same chain then the number value for the chain would be set to plural. In the same way a masculine singular pronoun could match to a chain having any value for person since the pronoun itself does not have a value. Substantives can be matched to referent chains already containing the same substantive or those which do not yet have a substantive.

When the system is run on 3 John using grammatical agreement only and selecting the nearest agreeing discourse referent, the system achieves an accuracy of seventy-eight percent. This is a high figure for just grammatical agreement and is due to the inflectional nature of Greek. This reduces the choices available to such a great degree that highly accurate figures can be achieved. The epistolary genre also helps in this regard as there is a clear distinction in this letter between the sender (1st person singular) and the received (2nd person singular).

2.1.3 Salience weights

The resolution algorithm uses the same basic structure for salience scores as the Lappin and Leass algorithm (see figure 3). A few changes have been necessary to include all the grammatical features of Greek. A value for implied references (those indicated by inflection) has been added with a weighting between those for the

Subject and Existential instances. The complements have also been split so that Dative compliments are given a slightly lower weighting than non-dative complements. This reflects the differing levels of grammatical involvement in the clause between direct object and indirect object complements. In addition the non-adverbial category has been removed. The figures currently used are shown in figure 4 but these can easily be adapted to optimise the algorithms performance.

Sentence Recency	100
Subject	80
Implied Reference	70
Existential	60
Complement (not Dative)	50
Complement (Dative)	40
Head Noun	80

Figure 4: Salience Factors used for Ancient Greek Anaphor Resolution.

When the Salience algorithm is run alongside the grammatical agreement filters the accuracy for tests on 3 John increase slightly to eighty-one percent. Due to the small size of the text this represents only two more correct resolutions, but it does give an insight into what features of language salience theory accounts for. An example is found in lines 12-14 in appendix 1. Here the pronoun in line 14 could have as its antecedent either 'the brothers' in line 12 or 'the strangers' in line 13. With grammatical agreement only 'the strangers' is incorrectly selected as this is closest to the pronoun. With the salience algorithm working the correct antecedent, 'the brothers' is selected. This is because, although both discourse referents receive the same salience weight from clauses 12 and 13, 'the brothers' have already appeared in the discourse back in line 6 and so the salience weight is higher.

2.1.4 Co-reference

Co-reference rules have proved to be the most complex area in developing the algorithm and there is still some work to do. The task is made easier in some respects with the high levels of annotation present in the corpus which allows very precise conditions to be specified. At the present time the co-reference part of the algorithm disallows the following:

- co-reference between head terms in the same clause (with the exception of reflexive pronouns and embedded clauses)
- co-reference between elements in the same word group (this causes problems for intensive use of pronoun)
- co-reference between the subject of a genitive absolute clause and the subject of the following clause

When applied to 3 John these co-reference rules actually reduce the level of accuracy to seventy-six percent which is slightly below that achieved by the grammatical filters alone. One of the problems caused is the intensive use of the

pronoun in line 32 where the co-reference rules prevent the correct assignment. The co-reference rules were actually developed using a sample of narrative text (Mark chapter 5) and do help improve accuracy in that text. Using the same rules with 3 John suggests, as has been shown in the development of anaphor resolution algorithms in English, that the algorithms perform best when optimised for specific genre. More analysis of the affect of the co-reference rules when used with the different texts in the corpus are needed before any firm conclusions can be drawn about their overall usefulness across genre.

3. Architecture and the interface

Although there are still areas of the algorithm that can be further improved the algorithm on its own will never be able to produce results that are guaranteed to be reliable. In addition there are some tasks with anaphora resolution that cannot be solved algorithmically such as distinguishing between two characters with the same name or combining instances where the same participant is referred to with different substantives. Also from the perspective of annotation not every discourse referent will be of interest as a participant so the final list must be editable. For these reasons human intervention is required in order to check and complete the participant annotation.

The algorithm described above forms the basis of a web application that enables a user to check and correct an assignment and then rerun the algorithm with the changes. For example the user may indicate a correction to a word assignment. Once this data is submitted and the algorithm rerun this word will be assigned as the user requested but also there could be changes further down the algorithm because of the changes in salience scores caused by the reassignment. The process then becomes an interactive one between the user and the algorithm.

When the user opens the web page the relevant xml files are loaded from the OpenText.org database, the anaphora resolution algorithm is run and the user is presented with an interface as shown in figure 5. The interface allows the user to highlight a participant chain providing a visual aid for checking the results. The user can then make a change to an assignment, this is done on a word by word basis with the incorrectly assigned word being indicated as a member of an alternative participant chain. These changes are stored in the DOM (Document Object Model) lying behind the interface. Once a change has been made the algorithm can be rerun. This may then fix incorrect assignments later in the document as it will affect salience scores and also potentially the information stored for gender, person and number for the corrected reference chain.

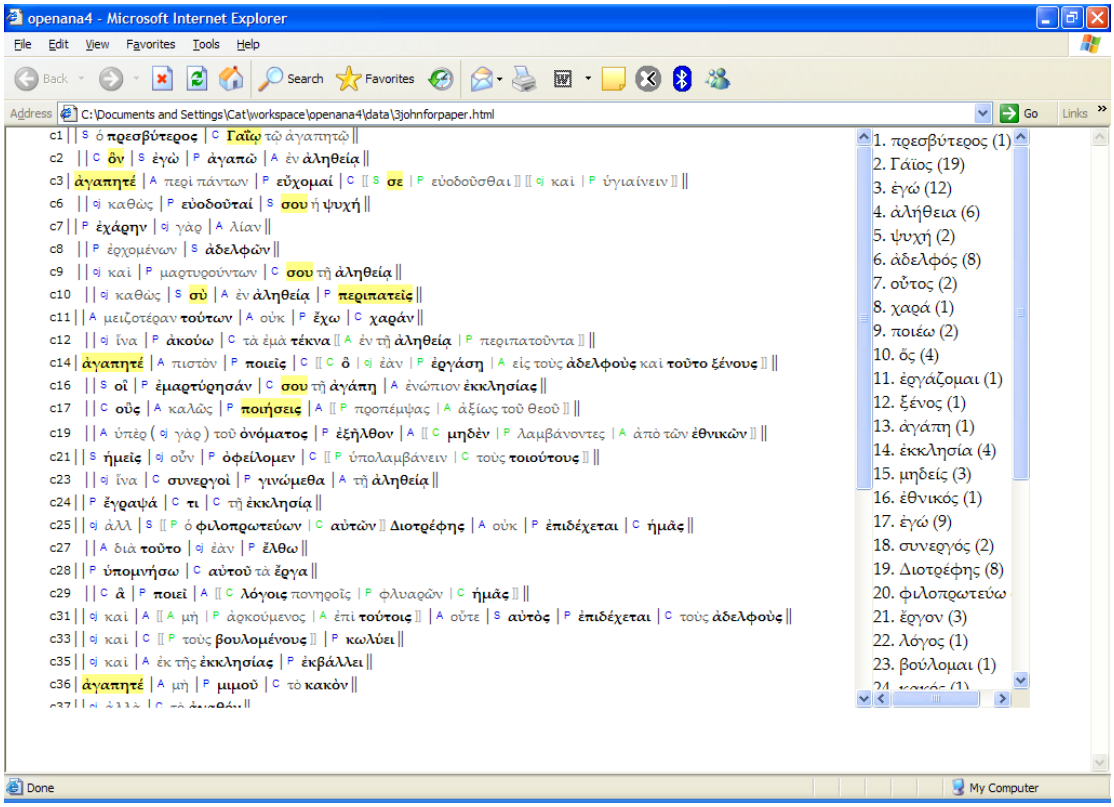

Figure 5: The user interface with one of the participant chains highlighted.

```
<participants>
    <wg.part num="1" href="NT.3Joh.w2" />
    <wg.part num="2" href="NT.3Joh.w3" />
    <wg.part num="2" href="NT.3Joh.w6" antecedentRef="NT.3Joh.w3" />
    <wg.part num="3" href="NT.3Joh.w7" />
    <wg.part num="3" href="NT.3Joh.w8" />
    <wg.part num="4" href="NT.3Joh.w10" />
    <wg.part num="2" href="NT.3Joh.w11" antecedentRef="NT.3Joh.w6" />
    <wg.part num="3" href="NT.3Joh.w14" />
    <wg.part num="5" href="NT.3Joh.w20" />
    <wg.part num="5" href="NT.3Joh.w23" />
    <wg.part num="3" href="NT.3Joh.w24" />
    <wg.part num="6" href="NT.3Joh.w28" />
    <wg.part num="4" href="NT.3Joh.w33" />
```

Figure 6: A sample section of the xml used to store the participant information.

The changes made by the user in a session are output to an xml file (see figure 6) which records the internally assigned participant number (the num attribute); a reference to the word being assigned to the participant (the href attribute) and if any word has been reassigned it also records the word number of the preceding word in the chain (the antecedentRef attribute). This xml is used by the algorithm and any user specified assignments override the algorithms internal choices thus correcting the previous error and potentially changing other decisions later in the text. Once the user is happy with the result or finishes an editing session the resulting xml is stored back to the OpenText.org database. The participant annotation can then be used as the basis

of a variety of different views and can be reloaded into the annotation interface if further changes need to be made.

4. Conclusion

Anaphora resolution is still an important research area within Natural Language Processing (NLP). Its importance comes, in part, from its nature as a low-level language feature. Any high-level processing task, such as machine translation and text summarisation, could be hugely improved if they were able to worth with a text having all of the pronouns correctly resolved to their antecedent. Even text retrieval tasks, such searching the internet, could be made more accurate and comprehensive with a reliable anaphora resolution system. In a similar way some of the questions asked of corpus data could also be more fully answered if such accurate anaphora resolution was available.

Mitkov *et al*. (2007) report that in order to start making a real difference to higher level tasks an accuracy level of at least eighty percent is required. For the tasks of interest to NLP research eighty percent accuracy may well be adequate enough but for work with corpus data this would still not suffice. Here the computer aided annotation tool could prove to be the way forward by speeding up the process of annotation while achieving the highest level of accuracy possible.

References

Brennan, S.E., Friedman, M.W., and Pollard, C. (1987) 'A Centering Approach to Pronouns' in *ACL-87*. ACL. pp. 155–62.

Deoskar, T. (2004) Techniques for Anaphora Resolution: A survey http://www.cs.cornell.edu/courses/cs674/2005sp/projects/tejaswini-deoskar.doc

Hobbs, J.R. (1977) *38 examples of Elusive Antecedents from Published Texts*. Research Report #77-2, August 1977. Department of Computer Science, City College, City University of New York. www.isi.edu/~hobbs/ElusiveAtecedents.pdf

Hobbs, J.R. (1978) 'Resolving Pronoun References' *Lingua*, 44. pp. 311–38. Reprinted in B.J. Grosz, K. Sparck Jones, B.L. Webber (eds.). 1986. *Readings in Natural Language Processing*. Los Atlos, CA: Morgan Kaufmann Publishers Inc. pp. 339–52. All page numbers in refer to the 1986 edition.

Jurafsky, D., and J.H. Martin (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. International Edition. London: Prentice Hall.

Lappin, S., and H.J. Leass (1994) 'An Algorithm for Pronominal Resolution' *Computational Linguistics*. vol. 2, no. 4. pp. 535–61.

Lappin, S., and M. McCord (1990a) 'A syntactic filter on pronominal anaphora in slot grammar'. in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. pp. 135–42.

Lappin, S., and M. McCord (1990b) 'Anaphora resolution in slot grammar' in *Computational Linguistics*, vol 16. pp. 197–212.

McCord, M. (1980) 'Slot Grammars' *American Journal of Computational Linguistics*. vol 6. pp. 31–43.

Mitkov, R. (1999) *Anaphora resolution: the state of the art*, Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.

Mitkov, R., R.J. Evans, C. Orasan, L.A. Ha and V. Pekar (2007) 'Anaphora resolution: To what extent does it help NLP applications?' in A. Branco *DAARC2007* LNAI 4410, Springer-Verlag, pp. 179–90 as presented at the Artificial Intelligence and Natural Computation seminar (School of Computer Science, University of Birmingham; 11 June 2007).

O'Donnell, M.B. (2005) *Corpus Linguistics and the Greek New Testament*. New Testament Monographs, 6; Sheffield: Sheffield Phoenix Press.

Okumura, M., and Tamura, K. (1996) 'Zero Pronoun Resolution in Japanese Discourse based on Centering Theory', in *Proceedings of COLING-96*. pp. 871-6.

Smith, C.J. (2005) *Casting out Demons and Sowing Seeds: A Fresh Approach to the* Synoptic Data from the Perspective of Systemic Functional Linguistics. University of Birmingham PhD thesis.

Tetreault, J.R. (1999) 'Analysis of Syntax-Based Pronoun Resolution Methods' in *ACL-99*. ACL. pp. 602–605.

Walker, M.A. (1989) 'Evaluating Discourse Processing Algorithms' in *ACL-98*. ACL. pp. 251–60.

Appendix I – A literal translation of 3 John arranged by clause

1. The Elder, to Gauis the beloved one
2. Whom I [I] love in truth
3. Beloved one! Before everything [I] pray that you are well and healthy
4. As [it] is well the soul of you
5. For [I] rejoiced greatly
6. When [they] came the brothers
7. And [they] gave witness to the truth of you
8. As you in truth [you] walk
9. Greater than these things not [I] have joy
   (I can have no greater joy than these things)
10. That [I might] hear that my children in the truth [they are] walking
11. Beloved one! Faithfully [you] do
12. the things [you] do for the brothers
13. and these strangers
14. who [they] gave witness to the love of you before the church
15. who [you will] do well sending in a manner worthy of God
16. because on behalf of the name [they] go out, receiving nothing from the gentiles
17. We therefore [we should] receive similar ones to these
18. so that fellow workers [we might] become in the truth
19. [I] wrote something to the church
20. but, the one wanting to be first of them, Diotrephes [he] would not receive us
21. because of this if [I] come
22. [I will] bring attention to the works of him
23. which [he] does with evil words slandering us
24. and not being satisfied with these things, nor does he [he]receive the brothers
25. and the ones wanting to [he] prevents
26. and out of the church [he] throws
27. Beloved one! Do not imitate the bad
28. but the good
29. the one doing good from God [he] is
30. the one doing bad [he] cannot see God
31. about Demetrius [it] is witnessed to by all
32. and by the truth itself
33. and we also [we] bear witness
34. and [you] know

35. that the witness of us [it] is true
36. much [I] have to write to you
37. but [I] do not wish with pen and ink to write to you
38. but [I] hope quickly to see you
39. and mouth to mouth [we will] speak
40. peace to you
41. [they] greet you the friends
42. greet the friends! Each by name