FACULTY OF HUMANITIES
UNIVERSITY OF COPENHAGEN

# Cover Page for Take-Home Assignments / Synopses

## For Assignments at the Faculty of Humanities

| Personal information | |
|---|---|
| Name | Gavin Patrick Abercrombie |
| Student email *(e.g. abc123@alumni.ku.dk)* | jst662@alumni.ku.dk |
| Phone number | 91968046 |

**Department**

- ○ ENGEROM
- ○ IKK
- ◉ INSS
- ○ IVA
- ○ MEF
- ○ SAXO
- ○ ToRS

| Study information | |
|---|---|
| Study programme | |
| Level | ○ BA ○ BA elective ◉ KA ○ MA elective ○ Open University |

| Information about the exam | | | | | |
|---|---|---|---|---|---|
| Title of course | Language Processing II | | | | |
| Title of exam | | | | | |
| Subject-element code | | | | | |
| Curriculum | 2013 | | | | |
| Examination period | Spring 2014 | | | | |
| Assignment title | Language Processing 2 - Task 2 | | | | |
| Standard pages | 15.5 | No. of characters | 31673 | May the assignment be incl. in the assignment library? | ◉ Yes ○ No |
| Examiner | Paggio, Patrizia (No. of characters includes 10116 technical) | | | | |

| Information in case of group examinations | | | | |
|---|---|---|---|---|
| The author of each section in the paper is clearly stated (tick off) | ◉ Yes | ○ No | | |
| Co-writer 1 | Name | Brian Philippi | KU username | prn776 |
| Co-writer 2 | Name | Thomas Andreas Sjögren | KU username | nzq844 |
| Co-writer 3 | Name | | KU username | |
| Co-writer 4 | Name | | KU username | |

### Declaration of Academic Honesty

I hereby confirm that I have completed this assignment on my own. In case of group examination, is the author of each section in the paper clearly stated in accordance with rules on group examinations. All quotations in the text have been marked as such and the assignment or fundamental parts of it have not been presented before in other contexts of assessment. The maximum number of standard pages has not been exceeded. *Handing in the assignment electronically by logging in at Absalon and uploading the material will replace a handwritten signature.*

You can find instructions on how to merge two PDFs on http://absalon.hum.ku.dk/english/students/onlinesubmission/

Rev. 23-April-2014

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

# Task 2

## Note of importance! (Thomas, Brian, Gavin)

**The report contains technical writing and programming code that should be handled as technical text when calculating number of pages.**

**The source code (.py file) is supposed to be run as it is with the Enthought version of Python. It requires nltk to have the Brown corpus installed (run `nltk.download()` in a python shell after importing nltk).**

## Part 1 (Gavin)

### Pronominal anaphora resolution

David Crystal's (2011) dictionary definition of **anaphora** is "the process or result of a linguistic unit deriving its interpretation from some previously expressed unit or meaning", and **anaphor** is described as the term used to label such a linguistic unit. Jurafsky & Martin (2009) explain **anaphora** in more simple terms: "Reference to an entity that has been previously introduced into the discourse is called anaphora, and the referring expression used is said to be **anaphoric**."

They also provide us with a definition of **reference resolution,** which they describe as being "the task of determining what entities are referred to by which linguistic expressions."

**Pronominal anaphora resolution** is therefore, the act of establishing which antecedents (previously mentioned entities) are referred to by which pronouns in a given text or language sample. In other words, it is a type of co-reference resolution in which the referring expression is a pronoun (Jurafsky & Martin, 2009).

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

## Phenomena accounted for by pronominal anaphora resolution

While the five basic types of anaphora listed by Jurafsky and Martin are *indefinite noun phrases, definite noun phrases, pronouns, demonstratives and names*, **pronominal anaphora resolution** obviously only accounts for the third of these phenomena, **pronouns**.

Pronouns usually refer to an entity that is previously mentioned as the antecedent of the pronoun, thus evoking that entity (Crystal, 2011). Take for example the following sentence:

> "The teachers prepared a fiendishly difficult exam. They thought it would challenge the students."

In this case a specific group of teachers is evoked by the **referent** *the teachers*, which is then referred to by the **referring entity**, or anaphor, the pronoun *they*.

## Problematic cases

Anaphora has attracted a great deal of attention both from linguists and philosophers (King, 2005).

Leaving aside philosophical questions related to semantic truth conditions, we are often faced with difficulties, when attempting anaphora resolution. Pronominal anaphora have several features which must agree with those of the entity to which they refer in order to make sense. An anaphor and its antecedent must agree in terms of *number*, *person*, *gender*, and *binding* or *syntactic constraints* (Jurafsky & Martin, 1999).

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

Problems arise when there is ambiguity as to which of two or more entities an anaphoric pronoun could refer to i.e. if there are multiple entities that fulfill these constraints. Jurafsky & Martin (1999) propose the following as factors to consider when resolving pronominal anaphora:

- *Recency* - the most recently mentioned entity is the most likely to be the referent.
- *Grammatical role* - subject position entities are more salient than those in object position.
- *Repeated mention* - entities that have been mentioned repeatedly are more salient than those mentioned once.
- *Parallelism* - preference can be produced by parallel sentence structures.
- *Verb semantics* - the meanings of some verbs influence the manner in which pronouns can be interpreted.

However, even taking into account these factors, problems can arise. For example in the following sentence:

> "Medicines can be harmful to young children. Make sure you keep them locked in the bathroom cabinet."

Here both *medicines* and *children* meet grammatical and syntactic constraints, and it is only world knowledge (depending on our attitudes to medicines and children) that determines which of these entities is the referent of *them* (American Psychological Society 2005). Indeed, without this world knowledge, other factors such as *recency* have us locking the children in the bathroom cabinet, rather than the medicines.

In certain cases a pronoun can refer to an entity that has not been previously mentioned, but to a *complement set*, which is merely implied. This is called *complement anaphora* and presents further problems for anaphora resolution, as the referent does not appear in the text or language sample (Nouwen, 2003 ).

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

## Part 2 (Gavin)

1. *John met Peter in the court one week ago.*
   Not applicable (there are no pronouns in this sentence).

2. *Ann met him in the parliament yesterday.*
   Pronoun: *him*
   Referent: *Peter* or *John*.
   *Peter* is preferred as it is the most recently mentioned entity, and it is parallel to the first sentence i.e. John met *Peter*, and Ann met *him*.

3. *She waved at him repeatedly.*
   Pronoun: *him*
4. Referent: *Peter* or *John*. Preference depends on resolution of 2, and is continued (due to repeated *mention* and *parallelism*) i.e. if Ann met Peter, then she waved at Peter, but if she met John, then she waved at John.

5. *That morning he was discussing with three English lobbyists.*
   Pronoun: *he*
   Referent: *Peter* or *John*. Again preference depends on resolution of 2, and continues (*repeated mention*).

   (N.B. This is an odd sentence as *discuss*, aside from literary and ancient usage, is a transitive verb which must take an object in both American English (Mirriam-Webster http://www.merriam-webster.com/dictionary/discuss) and British English (http://www.oed.com/view/Entry/54102?rskey=SVTmjn&result=2#eid)).

6. *They knew him very well and kept on talking.*
   Pronouns: *they* and *him*
   Referent: *three English lobbyists* and *Peter* or *John*. In the first case, *they* must refer to *three English lobbyists* due to feature agreement constrictions i.e. *three English lobbyists* is the only plural entity that can agree with *they*. The referent of *him* will be a continuation of the choice made in the previous sentence (*repeated mention*, *recency*).

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

7. *Peter did not notice Ann's waving.*

   Not applicable (no pronouns).

8. *She got upset about it.*

   Pronouns: *she* and *it*.

   Referents: *She* = *Ann* due to gender agreement, *it* = Peter's failure to notice Ann's waving due to agreement as the only inanimate entity, and to semantic verb agreement as the type of thing that people get upset about.

## Part 3 (Thomas, Brian, Gavin)

U0 John met Peter in the court one week ago.

1. a: referring expressions

   John Peter the court

   b: order by grammatical relation

   1. John (subject) 2. Peter (object) 3. the court ()

   c: sort and expand if pronominalized

   1. John (proper noun) 2. Peter (proper noun) 3. the court (noun)

   d: candidate list for backward-looking center (Cb)

   NIL, as this is the first utterance in the sequence, although, in Brennan, they just assume that the first utterance shown in the "discourse" is U+1 with the same Cp as their proposed Cb for their non-really-first utterance; clearly this is not consistent with the usual Centering theory idea that the Cb in a Continue transition should be pronominalized, in their U+1s ("Brennan," for example, in Figure 4) the Cb's are not pronominalized as

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

they would be had this not been their introduction. However, that is just a minor quibble with their method of exposition. It does nothing to invalidate their application of the theory per se, it merely suggests a potential weakness in it.

e: list proposed possible anchors

<NIL, ([JOHN:A1] [PETER:A2] [THE COURT:X1]*) >

*In very few of the examples in the research are inanimate nouns ever centered upon. Only in the case of the house, which is used as an example of an associative or bridging reference, does an inanimate object become the center of focus, and then probably only because the implications of using an animate (probably personal, if originating from the same authors) example are far too distracting, even potentially explosive. Of course, following that idea, it could very easily be presumed that "The court" could serve as a compelling center, if, as in the case of shock over the news that "the court" where John had just seen Peter has itself just been attacked with explosives, and any sort of imaginable intrigue between Peter and John could be used as an excuse for them being in the subject and object positions of the leading sentence. Of course, that would be a completely different story entirely. Still, the use case of this kind of algorithm is not as well explored in any of the literature.

2. a: filter by contraindices

b: filter by Cp agreement

c: filter by pronoun Cp match

Has to be NIL

3. a: rank proposed transitions

b: set values

<NIL, ([JOHN:A1] [PETER:A2] [THE COURT:X1]*) >

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

U1 Ann met him in the parliament.

1. a: Ann him the parliament

   b: Ann (s) him (obj) the parliament (obprp)

   c: him (John / Peter) (prn) Ann (pn) the parliament (noun)

   d: ([ANN:A3] [HIM/J:A1] [HIM/P:A2] [THE PARL:X2])

   e: <[ANN:A3], [ANN:A3] [HIM/J:A1] [HIM/P:A2] [THE PARL:X2]>

   <[HIM/J:A1], [ANN:A3] [HIM/J:A1] [HIM/P:A2] [THE PARL:X2]>

   <[HIM/P:A2], [ANN:A3] [HIM/J:A1] [HIM/P:A2] [THE PARL:X2]>

   <[PARL:X2], [ANN:A3] [HIM/J:A1] [HIM/P:A2] [THE PARL:X2] >

   <NIL, [ANN:A3] [HIM/JOHN:A1] [HIM/PETER:A2] [PARL:X2]

2. a:
   John and Peter are contraindexed with respect to the pronoun “him.” John takes precedence according to this implementation, because he occupied the subject position in U-1. Thus, anchor 3 from 1.e is eliminated

   b: Cb:him matches one anchor, namely 2 from 1.e.

   c: The proposed anchor is pronominalized, passing this filter

3. a: Shift

   b: Cb: him/John Cf: Ann, the parliament

It’s confusing how Brennan explains their algorithm in the text. It looks at just one utterance, but the Cf to Cb selection doesn’t make sense without both sides of the utterance to utterance transition in view. This is evidenced by the fact that the examples stop where there should be new ones coming from the finalizing utterance-one in the

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

sequence. Nonetheless, when this missing token is discovered, it makes the whole system activate, as it should.

U2 She waved at him repeatedly.

Continue

Cb: She:Ann

Cf: She:Ann, him:John

U3 That morning, he was discussing with three English lobbyists.

Retain

Cb: he:John

Cf: he: John, E. lobbyists

U4 They knew him very well and kept on talking.

Shift

Cb: They:E. lobbyists

Cf: They:E. lobbyists, him:John

U5 Peter did not notice Ann's waving.

Undefined

Cb: NIL

Cf: Peter, Ann's waving

U6 She got upset about it.

Shift

Cb: Ann (bridging reference: Ann's waving)

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

Cf: She:Ann, it:Peter's unawareness

## Are there any pronominal resolution preferences, which are not accounted for by the algorithm? (Thomas)

Indeed, the Brennan interpretation produces a certain preference for John over Peter, where that's not so clearly called for in the text. John, being the entity in the subject position that is eligible to be substituted by the pronoun "him" in the following phrase is given preference over Peter, in object position, even though, if the algorithm were to account for parallelism. This could be achieved by comparing the verb and its associated parameters, comparing its subject and object in U-1 and U0, which, finding continuity between the object entity, could override the usual grammatical role precedence of subject preferred over object as Cb U0. Of course, several sentences embedded into the text we find that Peter is the correct antecedent for him in all of the phrases, U1 through U4. There is no way for the CT implementation of Brennan to correct for this so far after the fact of the decision moment. Meanwhile, Jurafsky includes a discussion of ambiguity clarification in anaphora resolution systems, which hold multiple possible interpretations in play until one is definitely preferred in the text. Such feedback, in the form of a degree of certainty measure, could be incredibly meaningful to writing classification. It is conceivable that more difficult texts, such as Spencer's The Faerie Queene, or Pynchon's Gravity's Rainbow, could be shown to hold more potential antecedents in play and across longer spans of text, which require an acrobatic performance of the reader. This kind of text classification flows directly into the purview of Centering Theory, which seeks to clarify utterance sequence coherence and salience, just the elements artistically obscured by some of the world's most celebrated authors. On the other hand, the opposite output could be useful for texts where the asymptotically clearest message is called for, such as technical and medical manuals.

## Are all the pronouns in Text_a accounted by Centering? (Thomas)

Yes, all of the pronouns are accounted for, but they are not all clear and they are not all correct. Ann in the last utterance is the tricky case, because if she is taken to be the *Cb* of *U6* even though her reference in *U5* is not the head of its noun phrase, she is at least

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

explicited in *U5* more than the house whose door is ajar in the example of a bridging reference in Jurafsky and Poesio. According to the Brown Corpus tagger, "*Ann's*" would be a NP$, while according to the Penn Corpus tagger this would return NNP+POS: even though she is not the head of the noun phrase, Ann would be added to the *Cf* set. Secondly, and more importantly, the string of *he*s and *him*s that ought to refer to Peter but are preferred by the Brennan algorithm to resolve to John don't map to the text as is revealed in the *NIL Cb* and Undefined transition from *U4* to *U5*. These pronouns have a proper resolution in accordance with the process of the algorithm, but the inbuilt preference of the algorithm causes those pronouns to prefer the wrong antecedent. All of the pronouns are there, but their identification is not always correct.

## Part 4 (Thomas)

The solution we have created for this assignment is to take the sentences and tag each token with the help of the bigram-tagger included in NLTK, which we trained using the Brown corpus. The script then chunks it into sentence-trees that it uses to extract the noun phrases. Finally it takes out all the heads of the noun phrases, and appends to these agreement features. The script runs from start to end without any function call, but it can be divided into five distinct parts: Setup, tagging, chunking, extracting and appending.

### Setup

This first part of the script manages the setup of imports, texts, POS-tagger and grammar. The large part is reasonably simple but we want to pay additional attention to the setup of the POS-tagger.

We are using a series of taggers linked together with the `backoff` attribute. Appending a backoff-tagger to the instantiation of the class means that if a tagger has problems with assigning a tag to a token, it consults the provided backoff-tagger for help. In our case we start with setting up the `nltk.DefaultTagger()` to assign 'NN' to every token it encounters. The script is not tagging with this instance but uses it as consult, as described above. The next step is to set up the `nltk.UnigramTagger()` and provide the

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

`DefaultTagger()` as backoff. We are also providing the tagged corpus for this instantiation to train the tagger. The last and principal tagger is the `nltk.BigramTagger()` that we provide with the `UnigramTagger()` as backoff and the training set and save the instance to be used when assigning POS-tags to our sentences.

### Tagging

After the setup, the script starts iterating the sentences and the first step is to tokenize the sentences into a list of tokens that can be used in the POS-tagger. We are using `nltk.word_tokenize()` which provides us with a convenient way to accomplish this and forward the tokenized sentences to our POS-tagger.

The tagger then iterates through the tokens as bigrams to find the most common likely tag for each token based on the Brown corpus we used to train the taggers. As it's a bigram-tagger it uses the preceding tag when assigning. Note that Brown uses a different tag-set then Penn Treebank.

### Chunking

Next we are apply `nltk.RegexpParser()` which uses our previously defined grammar to chunk the tagged tokens and identifies the noun phrases (NPs), as that is what we are interested in. Our grammar is based on the texts we are to process but could easily be extended to cover larger variations of texts.

### Extracting

When we have the chunk-trees with the identified phrases, the script starts to extract the NPs and strips down to only storing the head nouns. The logic we have used for this is that compound nouns are built up of a describing word and the noun that gives the compound meaning. E.g. 'bicycle stand' where the compound is an extension of the word 'stand' while the word 'bicycle' only describes what type of stand it is.

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

## Appending

The final part of the script checks the tokens against the agreement feature dictionary. It is a simple checkup appending the matching feature to the end of the token list.

## Source code of the program (Thomas, Brian, Gavin)

```
# Imports
import nltk
from nltk.corpus import brown

# Load in the texts and split into sentences stored in a list.
texts = "John met Peter in the court one week ago.\nAnn met him in the parliament yesterday.\nShe waved at him
repeatedly.\nThat morning, Peter was discussing with three english lobbyists.\nHe did not notice his friend."
texts = texts.split("\n")

# Load training corpus from the Brown corpus. We are using all data available.
tagged = brown.tagged_sents()

"""
 Train the tagger, starting with Default tagger up to BigramTagger.
 This is the most time consuming part(!).
"""
t0 = nltk.DefaultTagger('NN')
t1 = nltk.UnigramTagger(tagged, backoff=t0)
t2 = nltk.BigramTagger(tagged, backoff=t1)

# Create a dictionary linking words to their agreement feature.
masc_sing = ['he', 'him', 'john', 'peter', 'zebediah']
fem_sing = ['she', 'ann', 'friend', 'margreth', 'zelda']
inam_sing = ['it', 'car', 'court', 'morning', 'english', 'notice', 'parliament', 'week', 'yesterday',
'xylophone']
plural = ['they', 'cars', 'lobbyists']

# Grammar to use for chunking.
grammar = r"""
 NP: {<DT>?<JJ>*<NP><NNS>?}
   {<DT>?<JJ>*<NP>|<NNS>}
   {<DT>?<JJ>*<NNS?><NNS?>?}
   {<NNP>+}
   {<NP>}
   {<NR>}
   {<PPS>+}
   {<PP$>}
   {<PPO>}
"""

chunk_parser = nltk.RegexpParser(grammar)

results = []

# Run the sentences.
```

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

```
# Aiming for the format of ([ 'john', 'NNP', 'masc_sing'],[ 'ann', 'NNP', 'fem_sing'],[ 'him', 'PRP',
'masc_sing'])
for text in texts:

 # Tokenize the sentence.
 tmp = nltk.word_tokenize(text)

 # Use the trained BigramTagger to assign Part-of-Speech tags to all tokens.
 tmp = t2.tag(tmp)

 # Chunking the sentences into trees so that we can extract NPs
 tmp = chunk_parser.parse(tmp)

 """
   Extract all elements from the chunked trees that are Noun Phrases.
   Sanitize the trees and only stores the head nouns. Removing delimiters, adjectives and so on.
 """
 np_list = []
 for elem in tmp:
   if isinstance(elem, nltk.tree.Tree) and elem.node == "NP":
     if len(elem) < 2:
       np_list.append(list(elem[0]))
     else:
       np_list.append(list(elem[1]))

 for i in np_list:
   i[0] = i[0].lower()

 # Add agreement features to the head nouns.
 for pair in np_list:
   if pair[0] in masc_sing:
     pair.append('masc_sing')
   elif pair[0] in fem_sing:
     pair.append('fem_sing')
   elif pair[0] in inam_sing:
     pair.append('inam_sing')
   elif pair[0] in plural:
     pair.append('plural')
   else:
     # If the word doesn't occur in the dictionary.
     pair.append('UNDEFINED')

 results.append(np_list)

for i in results:
 print tuple(i), '\n'
```

## Program output

```
(['john', 'NP', 'masc_sing'], ['peter', 'NP', 'masc_sing'], ['court', 'NN', 'inam_sing'], ['week', 'NN',
'inam_sing'])
(['ann', 'NP', 'fem_sing'], ['him', 'PPO', 'masc_sing'], ['parliament', 'NN', 'inam_sing'], ['yesterday', 'NR',
'inam_sing'])
(['she', 'PPS', 'fem_sing'], ['him', 'PPO', 'masc_sing'])
(['morning', 'NN', 'inam_sing'], ['peter', 'NP', 'masc_sing'], ['lobbyists', 'NN', 'plural'])
```

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

```
(['he', 'PPS', 'masc_sing'], ['friend', 'NN', 'fem_sing'])
```

## Part 5 (Thomas, Brian, Gavin)

To set up the function for finding the transitions between the parameter-sets provided, we have assumed that they are based on the sentences (Text_b) in part four. Based on this assumption we found all centers (forward, preferred and backwards) for each sentence as displayed in the *dictionary of centers* in the source code. This is important for linking the pronouns together with their referent nouns in previous sentences.

The function takes one parameter-set together with a counter, to match the set with the correct centers in our dictionary. It then look up and replaces pronouns with their referent nouns so that 'him' is equal to 'john' in the coming tests.

As provided in the assignment, the parameters need to match in token and agreement feature with the additional requirement that one must be a pronoun, following the rules from Jurafsky (2009) and Brennan (1987). The function assigns a score that decides if it's to be treated as in the first or the second column of the transition matrix and the following test assigns the correct transition to the utterance.

### Source code of the program

```
# The dictionary of centers for each utterance.
rules = {
 'u1': {
   'cf': ['john', 'peter', 'court', 'week'],
   'cp': ['john'],
   'cb': ['undefined']
 },
 'u2': {
   'cf': ['ann', 'him', 'parliament'],
   'cp': ['ann'],
   'cb': ['ann'],
   'link': {'him': 'john'}
 },
 'u3': {
   'cf': ['she', 'him'],
   'cp': ['she'],
   'cb': ['she'],
   'link': {'she': 'ann', 'him': 'john'}
 },
 'u4': {
   'cf': ['peter', 'lobbyists', 'morning'],
   'cp': ['peter'],
```

```
   'cb': ['undefined']
 },
 'u5': {
   'cf': ['he', 'friend'],
   'cp': ['he'],
   'cb': ['he'],
   'link': {'he': 'peter', 'friend': 'ann'}
 }
}

# The parameter sets provided in the assignment.
ut = [
 (['john', 'NNP', 'masc_sing'],['ann', 'NNP', 'fem_sing'],['him', 'PRP', 'masc_sing']),
 (['him', 'PRP', 'masc_sing'],['she', 'PRP', 'fem_sing'],['she', 'PRP', 'fem_sing']),
 (['she', 'PRP', 'fem_sing'],['morning', 'NNP', 'anim_sign'],['peter', 'NNP', 'masc_sing']),
 (['peter', 'NNP', 'masc_sing'],['he', 'PRP', 'masc_sing'],['he', 'PRP', 'masc_sing'])
]

"""
 The function that is calculating transitions between the utterances.

 Taking the Cb(Un), Cp(Un+1), Cb(Un+1) and the counter as parameters.

 Printing the calculated transitions to the screen.
"""
def transition(cb1, cp2, cb2, count):

 # Check for PRPs in the parameters and link them together with their corresponding part of previous utterance.
 # This results in that 'him' in parameter set two is linked together with 'john' from the first utterance.
 if cb1[1] == 'PRP':
   cb1[0] = rules['u' + str(count)]['link'][cb1[0]]

 if cp2[1] == 'PRP':
   cp2[0] = rules['u' + str(count + 1)]['link'][cp2[0]]

 if cb2[1] == 'PRP':
   cb2[0] = rules['u' + str(count + 1)]['link'][cb2[0]]

 """
   Scoring the transitions based on the transition rule-set.
   Score or 1 in the first test means that it is in the first
   column of the transition matrix. Score of 2 is in the second column.

   We are following the equality rule provided in the assignment
   by checking if tokens and agreement features are the same. In case of true
   we check if one is a pronoun.
 """
 score = 0
 if cb2[0] == cp2[0] and cb2[2] == cp2[2]:
   if cb2[1] == 'PRP' or cp2[1] == 'PRP':
     score = 1
 else:
   score = 2

 # The second part of the scoring is giving us the actual transition.
 # Instead of columns we are now testing for the rows.
```

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

```
 if score == 1:
   if cb2[0] == cb1[0] and cb2[2] == cb1[2]:
     if cb2[1] == 'PRP' or cb1[1] == 'PRP':
       trans = 'Continue' # 11
   else:
     trans = 'Retain' # 12

 if score == 2:
   if cb2[0] == cb1[0] and cb2[2] == cb1[2]:
     if cb2[1] == 'PRP' or cb1[1] == 'PRP':
       trans = 'Smooth-shift' # 21
   else:
     trans = 'Rough-shift' # 22

 # Print out the results.
 print trans

# Use a counter to keep track of what rule to refer to.
count = 1

# Iterate the parameter sets and run the function.
for u in ut:
 transition(u[0], u[1], u[2], count)
 count += 1
```

## Program output

```
Smooth-shift
Retain
Rough-shift
Continue
```

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

## Part 6 (Brian)

**a)**

Roughly, centering transitions illustrate the development of topicality in a discourse by modeling entity salience and coherence through a sequence of utterances (although, most of the examples in the research involve monologues as opposed to interactive discourses, as Jurafsky points out). After the approach was popularized by Grosz et al. (1995) in the early half of the 1980's, several various implementations were developed by subsequent researchers. The history of this development is carefully catalogued by Poesio et al. (2005) as the background in which they aim to put these different specific varieties of Centering Theory into perspective by testing their performance computationally.

Grosz et al. (1995) identify three kinds of transitions, *Continue*, *Retain*, and *Shift*, in increasing order of inferential load value. A *Continue* transition has the same *Cb* in the next *Utterance* as the previous one, where this *Cb* is also the *Cp*: it is predicted to be the *Cb* in the next. A Retain transition has the same *Cb* in the next *Utterance* as the previous one, but the *Cp* of the previous *Utterance* does not match the *Cb* in the next one. A *Shift* transition has a different *Cb* than the previous *Utterance*, and while Grosz et al. don't distinguish types of *Shift* transitions, Brennan et al. introduce a refinement: a *Smooth Shift*, where the *Cb* in the next *Utterance* matches the *Cp* of the previous, and a *Rough Shift*, where the *Cb* in the next and the *Cp* in the previous *Utterance* do not match. With this added distinction, the increasing order of inferential load transition types yields: *Continue*, *Retain*, *Smooth Shift*, and *Rough Shift*, with the earliest shift possible preferred because it minimizes inferential load, or in other words maximizes coherence and aligns salient features.

**b)**

Pronominal resolution algorithms attempt to attribute pronouns to the correct antecedent without recourse to a discourse model, relying instead on semantics and

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

exception-rule preferences to pinpoint the correct pronoun for a given anaphoric entity. Centering theory, like Lappin and Leass' algorithm, instead models discourse. What distinguishes CT from LL is that CT does not weight references with a points system like LL, and it claims that there is only one center at a time. Further, CT focuses on categorizing the degree of discourse coherence by tracking transitions from one Utterance to the next, as discussed above.

**c)**

According to Jurafsky & Martin (2009, p.688):

> *"the centering algorithm was developed on the assumption that correct syntactic structures are available as input. In order to perform an automatic evaluation on naturally occurring data, the centering algorithm would have to be specified in greater detail".*

We therefore encounter a variety of problems when applying the theory to texts, and these problems can be lessened or magnified, depending on the choices made in application. Here we list the major difficulties:

**Hyper-Programmatic**

Centering Theory proceeds with fairly rigid fidelity to a programmatic approach to entity tracing, which leads to awkwardness in its discourse model. This is clear from its example sentences, which don't usually exemplify natural language, and from its mishandling of sentences that do better represent such language. In CT, the preference for programmatic coherence clashes with the more fluid means of entity establishment and elaboration employed in natural language. Grosz et al. (1995) identify this shortcoming, "the lack of complete information," which is the primary trigger for confusion in the model: CT presumes complete information and penalizes deviations from that kind of procession through the discourse. Grosz et al. also identify implementation improvements with possibly better fluidity in this respect, but purely in

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

a conjectural manner, without experimental results to determine which of the possible improvements are in fact promising developments.

**EST: establishment**

Kameyama (1986) proposed an EST:establishment feature to alleviate some of the tension felt when CT suddenly finds no Cb for a given utterance. Her suggestion hasn't taken much root, but it is interesting to note, since the lack of a fitting center is disruptive to the model's results and frustrating from a researchers perspective. Furthermore, the notion of an establishing center maps well to the intuitive expectation of a human interpreter.

**Cheap vs. Expensive Transitions**

Strube and Hahn (1999) highlight, not necessarily a fault in the CT as it is at present, but an area of application that has been little developed or applied: *cheap* vs. *expensive* transitions. It's a fortunate wording compared to the smooth / rough distinction to which it is quite similar, but it emphasizes Strube and Hahn's point that CT's strength is in its claim to predict the Cb via the Cp. Using the idea of cheap versus expensive allows for the strengths of both kinds of transitions to be embraced and contrasted, versus narrowly preferred, as is the case in the Shift and Smooth vs. Rough Shift denominations. This allows room for expansion into various kinds of texts, and indeed, application to cover more demanding texts, which in some cases are good or even necessary. In this way, however, CT could be expanded to classify when and potentially exactly how a text is more or less demanding of the reader, which of course can also be useful for the interpretation of the text.

**Problems arising in naturally occurring language**

In naturally occurring language, there are often no Backward-looking Centers. If fact Poesio et al. (2004) found that in some domains as little as 10% of utterances contained one.

In fact the domain of a text has a large bearing on how effectively Centering theory can be applied to it. For example, in pharmaceutical leaflets, amongst other texts, it is

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

common to find pronouns, which refer to a global center, an entity which is the main subject of the text and is mentioned frequently, particularly at the start. These global centers are then preferred over other, more local options as the referent of those pronouns (Poesio et al., 2004).

The way in which a text is segmented is also an issue when applying Centering theory, and whether or not to treat relative clauses as separate utterances, and more generally, how to define utterance, segmentation and paragraphs in relation to Centers, can lead to variance in performance depending on the text (Poesio et al., 2004)

Thomas Sjögren – nzq844
Brian Philippi – prn776
Gavin Abercrombie – jst662

# References

American Psychological Society. (2005, August 8). Which It Is It? Resolving Ambiguous Pronouns: English Teachers Would Be Proud. ScienceDaily. Retrieved June 7, 2014 from www.sciencedaily.com/releases/2005/08/050805103610.htm

Crystal, David. Dictionary of linguistics and phonetics. Vol. 30. John Wiley & Sons, 2011.

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. (1987). A centering approach to pronouns. In Proceedings of the 25th Meeting of the Association for Computational Linguistics, pages 155-162, Stanford, CA. http://dl.acm.org/citation.cfm?id=981197

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. Computational linguistics, 21(2), 203-225.

D. Jurafsky and J. Martin (2009). Speech and Language Processing. Prentice-Hall

Kameyama, Megumi. 1986. A property-sharing constraint in centering. In Proceedings of the ACL-86, New York, pages 200–206.

King, Jeffrey C., "Anaphora", The Stanford Encyclopedia of Philosophy (Summer 2013 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/sum2013/entries/anaphora/.

Nouwen, R. (2003). Complement anaphora and interpretation. Journal of Semantics, 20(1), 73-113.

Poesio, M., Stevenson, R., Di Eugenio, B., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. Computational linguistics, 30(3), 309-363.

Pynchon, Thomas. Gravity's Rainbow. Viking Press, 1973.

Spenser, Edmund. The Faerie Queene. Penguin Classics, 1979.

Strube, Michael and Udo Hahn. 1999. Functional centering–Grounding referential coherence in information structure. Computational Linguistics, 25(3): 309–344.