A Predictive Model for Notional Anaphora in English

Amir Zeldes
Department of Linguistics
Georgetown University
Washington, DC, USA
amir.zeldes@georgetown.edu

Abstract

Notional anaphors are pronouns which disagree with their antecedents' grammatical categories for notional reasons, such as plural to singular agreement in: "the government ... they". Since such cases are rare and conflict with evidence from strictly agreeing cases ("the government ... it"), they present a substantial challenge to both coreference resolution and referring expression generation. Using the OntoNotes corpus, this paper takes an ensemble approach to predicting English notional anaphora in context on the basis of the largest empirical data to date. In addition to state of the art prediction accuracy, the results suggest that theoretical approaches positing a plural construal at the antecedent's utterance are insufficient, and that circumstances at the anaphor's utterance location, as well as global factors such as genre, have a strong effect on the choice of referring expression.

1 Introduction

In notional agreement, nouns which ostensibly belong to one agreement category are referred back to using a different category, as in (1) (Quirk et al., 1985), with singular/plural verb and pronoun.

(1) [The government] has/have voted and [it] has/[they] have announced the decision

Although examples such as (1) are often taken to represent a single phenomenon, subject-verb (SV) agreement and pronoun number represent distinct agreement phenomena and can disagree in some cases, as shown in (2) and (3), taken from the OntoNotes corpus (Hovy et al., 2006).

(2) [CNN] **is** my wire service; [they]**'re** on top of everything.

(3) [One hospital] in Ramallah **tells** us [they] **have** treated seven people

While previous studies have focused on SV agreement (den Dikken 2001, Depraetere 2003, Martinez-Insua and Palacios-Martinez 2003), there have been few corpus studies of notional pronouns, due at least in part to the lack of sizable corpora reliably annotated for coreference, and the low accuracy of automatic systems on difficult cases. In this paper we take advantage of the OntoNotes corpus, the largest corpus manually annotated for coreference in English (about 1.59 million tokens with coreference annotations), to build a predictive model of the phenomenon, which can be used for both coreference resolution and referring expression generation (see Krahmer and van Deemter 2012 for an overview).

2 Previous work

Theoretical linguistic discussions have focused on SV agreement, especially in expletive constructions (ECs, Sobin 1997; i.e. 'there is' vs. 'there are'). Reid (1991) discusses SV agreement and notional pronouns, and posits reference to persons as facilitating plural pronouns, as in (4) and (5), where a relative 'who' forces a +PERSON reading.

(4) And this fall [the couple] expects [its] first child.

(5) A Florida court ruled against [a Pennsylvania couple] **who** contend May's 10-year-old daughter is actually [their] child.

This suggests that inferred entity type may be a relevant predictor of notional anaphora. Other theoretical papers suggest a formal analysis with empty pronoun heads bearing a plural feature, e.g. the analysis in (6) from den Dikken (2001) (see also Sauerland 2003 for a similar analysis).

(6) $[_{\text{DP1}}$ pro$_{[+\text{pl}]}$ $[_{\text{DP2}}$ the committee$_{[-\text{pl}]}]]$ are ...

Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference, pages 34–43
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

This suggests that speakers decide on the notional agreement category already at the point of uttering the antecedent. However psycholinguistic studies have shown effects localized to the point of uttering the anaphor, due to processing constrains (see Eberhard et al. 2005, Wagers et al. 2009, Staub 2009). We hypothesize that processing constraints may make it difficult for speakers to remember the exact expression used for the antecedent after a long distance from the first point of utterance, and therefore consider some length and distance-based metrics as features below (see Section 3.3).

Corpus-based studies have shown that notional anaphora likelihood varies by modality (more often in speech), variety of English (more often in UK English) and genre (see Quirk et al. 1985: 758, Leech and Svartvik 2002: 201, Levin 2001). Depraetere (2003) explored the idea that verb semantics influence agreement choice, especially whether verbs imply decomposition or categorization of the unit (e.g. *consist of, be gathered, scatter*), or signify differentiation within a set (e.g. *disagree, quarrel*). Annala (2008) provides a detailed corpus study of nine nouns in the written part of the British National Corpus (`http://www.natcorp.ox.ac.uk/`) and the Corpus of Late Modern English Texts (CLMETEV, `http://www.helsinki.fi/varieng/CoRD/corpora/CLMETEV/`). The study found tense to be relevant for the nine nouns, with past tense of 'be' being particularly susceptible to triggering plural agreement, while for nouns which generally prefer plural, singular agreement appeared more often in the present. Taken together, these studies suggest that tense and verb classes may be relevant features, as well as indicating the importance of some conventional usage effects. The latter are also backed by psycholinguistic evidence that speakers process notional anaphora more quickly than strict agreement in contexts that are biased towards the non-agreeing plural (Gernsbacher, 1986).

3 Experimental setup

3.1 Types of cases included

In this paper we focus exclusively on plural pronouns referring back to singular headed phrases, but the exact nature of cases included requires some decisions. Since the number for second person pronouns (*you*, *your*, etc.) is ambiguous, we omit all second person cases. First person cases are rare but possible, especially in reference to organizations, as in (7), taken from OntoNotes.

(7) Bear Stearns stepped out of line in the view of [the SEC]$_i$ ... [we]$_i$'re deadly serious about bringing reform into the marketplace

Some of the same lexical heads can appear with both singular and plural first person reference, either for metonymical reasons ("when a country says 'I/we'...") or by coincidental homonymy.[1] These cases are therefore all included whenever a relevant NP is annotated as coreferent in OntoNotes.

Three main types of plural reference to singular antecedents can be distinguished in our data (see Section 3.2 for some statistics): the most common, which will be referred to as Type I, is reference to complex/distributive entities (so called 'committee' nouns) seen e.g. in (2). These are distinct from Type II, which has bleached quantity noun heads, (e.g. 'a number of X' or 'a majority of X') which may sometimes be referenced as a plurality, as in (8), and sometimes as a unit, as in (9).

(8) [the vast silent **majority** of these Moslems] are not part of the terror and the incitement , but [they] also do not stand up political leaders

(9) [the vaunted Republican **majority**] is just not now nor has [it] ever been ready for prime time governing

A third type (Type III) occurs in cases such as (10), denoting unspecified gender (these are sometimes called generic or epicene pronouns; see also Huddleston and Pullum 2002:493-494, Curzan 2003). This construction has been gaining popularity (Paterson, 2011), and has recently been approved by the 2017 Associated Press Stylebook as standard (`https://www.apstylebook.com/`).

(10) I'll go and talk to [the person here] cause [they] get cheap tickets

Although this type of agreement is semantically and pragmatically very different from the other two types above, it must be addressed in this paper for several reasons. Firstly, if we want to be

[1] For example one speaker in a forum discussion in the corpus has the user name 'A Very Ordinary Native Country', leading to coreference with the pronoun 'I'.

able to predict pronoun form for computational applications such as coreference resolution or natural language generation, then such cases should be covered in some way. Secondly, there are cases in which either a computer, or in some cases even a human would find it difficult or impossible to be sure of the class that a case falls under, as shown in (11) and (12), both real examples from OntoNotes.

(11) [The enemy] attacked several times, and each time [they] were repelled

(12) [a publisher] is interested in my personal ad book ... I looked [them] up

While in (11) it may seem unlikely that use of 'they' is meant to obscure gender, this reading cannot be ruled out, especially by automatic analysis. In (12), it is possible to get either reading: either the 'publisher' is a company, and therefore plural (Type I; but notice singular 'is' as a verb), or the speaker spoke with the director of a publishing house, disregarding that person's gender (Type III). Note that in singular agreement, these would result in saying 'she' or 'he' versus 'it', as in (13) (also from OntoNotes).

(13) [the Des Moines-based publisher] said [it] created a new Custom Marketing

Additionally, there are Type III cases in which plural pronoun agreement for singular-like reference is not motivated by gender constraints, e.g. (14).

(14) [Nobody] is going to like Bolton a year from now, are [they]?

Due to these complications, we include all cases of plural anaphora annotated as coreferent with singular NPs, though we will re-examine these types in the data in analyzing the results.

3.2 Data

The data for the present study comes from the OntoNotes corpus (Hovy et al., 2006), Version 5, the largest existing corpus with coreference annotations. OntoNotes contains gold POS tags and syntactic constituent parses, as well as coreference resolution for pronominal anaphora and definite or proper noun NPs (but not for indefinites, see below), and named entity annotations for proper nouns. The coreference annotated portion of the corpus contains 1.59 million tokens from multiple genres, presented in Table 1.

Table 1: Coarse text types in OntoNotes

Spoken		**Written**	
bc.conv	137,223	news	68,6455
bc.news	244,425	bible	243,040
phone	110,132	trans.	98,143
		web	71467
total	491,780	**total**	1,099,105
total 1,590,885			

Written data constitutes the large bulk of material, primarily from newswire (Wall Street Journal data), as well as some data from the Web and the New Testament, and some translations of news and online discussions in Arabic and Chinese. The translated data has been placed in its own category: it behaves more conservatively in preferring strict agreement than non-translated language (see Section 4.2), perhaps due to translators' editorial practices. The spoken data comes primarily from television broadcasts, including dialogue data from MSNBC, Phoenix and other broadcast sources (bc.conv), or news, from CNN, ABC and others (bc.news), as well as phone conversations.

The relevant cases from the corpus for the present study were extracted by finding all lexical NPs headed by singulars (tagged NNP or NN) whose phrases are referred back to by an immediate antecedent (the next mention) which is a first or third person pronoun, then filtering to keep only those singular NPs headed by a token which is attested as taking plural agreement somewhere in the corpus, but also including its attestation with singular pronouns. In other words, this study makes no *a priori* interpretation of anaphora as notional in isolation: all and only items actually attested in both forms are considered.

These selection criteria, followed by manual filtering for errors, led to the extraction of 3,488 anaphor-antecedent pairs, of which 1,209 exhibited notional agreement (34.6%), including a subset of 207 cases (5.9% of the data) which were unambiguously identifiable as Type III, gender neutral plural pronouns.

OntoNotes contains 17,263 direct anaphoric links to a singular NP, meaning we can estimate the frequency of all agreement types addressed here at a not insubstantial 7% of pronominal reference to a singular lexical NP antecedent, with gender neutral type III at about 1.2% and Types I-II covering 5.8% of the total corpus.

As a test data set, we reserve a random 10% of the data, amounting to 349 cases, stratified to include approximately the same proportions of genres, as well as notional vs. strict agreement cases. This stratification is important in order to test the classifier in Section 4.1 using realistically distributed data.

3.3 Feature extraction

To predict the occurrence of notional anaphora we will use a range of categorical features indicated to be relevant in previous studies (see Section 2): POS tags and dependency functions for the anaphor, antecedent and their governing token, entity types, genre/modality, and definiteness/previous mention of the antecedent. These features indirectly give us access to tense, grammatical constructions and some measures of salience (especially subjecthood and repeated mention). Additionally, we will consider a number of numerical features which may be relevant from a processing perspective, such as the distance in tokens between the anaphor and antecedent, length in characters and tokens for the antecedent NP, document token count, and the positions of the expressions in the document, expressed as a percentage of document length (e.g. an antecedent may begin at the 75th percentile of document token count). Most of these features can be extracted from the data automatically.

A limitation of using OntoNotes is that many antecedents of pronominal anaphora are not named entities (unnamed 'committees', etc.), meaning we do not have gold entity types for all NPs. In order to overcome this problem and expand the range of features available in this study, the entire corpus was annotated automatically for non-named entities using xrenner, a non-named entity recognizer (Zeldes and Zhang, 2016).

A second problem is that the coreference annotation guidelines for OntoNotes preclude antecedents for indefinite NPs, meaning cases such as (15) are marked as multiple entities (BBN Technologies 2007:4).

(15) [Parents]$_x$ should be involved with their children's education at home, not in school. [They]$_x$ should see to it ... [Parents]$_y$ are too likely to blame schools for the educational limitations of [their]$_y$ children.

The second instance of 'parents' is regarded as a separate, 'discourse new' entity. This will be relevant for using previous mention of the antecedent as a feature: we can only detect previous mention of the antecedent if it is annotated, and this will never be the case for indefinites.

In order to assess the influence of grammatical function and semantic classes of verbs governing either the anaphor or the antecedent, the syntax trees in the corpus were converted to a dependency representation using Stanford CoreNLP (Manning et al., 2014), allowing for a simpler use of dependency functions as a predictor. This also allows us to identify the governing verb (or noun etc.) for each mention. Governing verbs were then tagged automatically using VerbNet classes (Kipper et al. 2006), which give rough classes based on semantics and alternation behaviors in English, such as ALLOW for verbs like {*allow, permit, ...*} or HELP: {*aid, assist, ...*}, etc.

Because some verb classes are small or rare, potentially leading to very sparsely attested feature values, classes attested fewer than 60 times were collapsed into a class OTHER (for example VerbNet class 22.2, AMALGAMATE). Verbs attested in multiple classes were always given the majority class, for example the verb *say* appears both in VerbNet class 37.7 SAY and class 78, INDICATE, but was always classified as the more common SAY. Finally, some similar classes were collapsed in order to avoid replacement by OTHER, such as LONG + WANT + WISH, which were fused into a class DESIRE. Nominal governors (e.g. for possessive NPs, whose governor is the possessor NP) were classified by their NER entity class or non-named class predicted by the entity recognizer.

4 Results

4.1 Predictive ensemble model

In this section we construct a model to predict, given properties of a singular antecedent NP from a lexeme known to exhibit notional agreement, and properties of the position of the anaphor referring back to it, whether or not the pronoun will in fact be plural. Considering the highly contextual nature of notional anaphora, we would ideally want to use the entire sequence of text before and after each of the entity mentions to predict the choice of pronoun, for example using a Recurrent Neural Network. However, despite being the largest available dataset for English, the amount of gold standard examples we have (less than 4,000) makes a Deep Learning approach problematic. We therefore train an ensemble of decision trees on the features presented in Section 3, more specifically using the Extra Trees algorithm (Geurts et al., 2006), which outperforms the standard Random Forest algorithm and linear models on our data.

Using a grid search with 5 fold cross validation on the training data, the optimal hyper-parameters for the classifier were found, leading to the use of 300 trees with unlimited depth, limited to the default number of features in the scikit-learn implementation, which is the square root of the number of features rounded up. The best performance was achieved using the 20 features outlined in Figure 1, meaning that each tree receives 5 features to work with, thereby reducing the chance of overfitting training data. The classifier achieves a classification accuracy of 86.81% in predicting the correct form in the test set, an improvement of over 20% above the majority baseline of always guessing 'strict agreement' (65.6% accuracy).

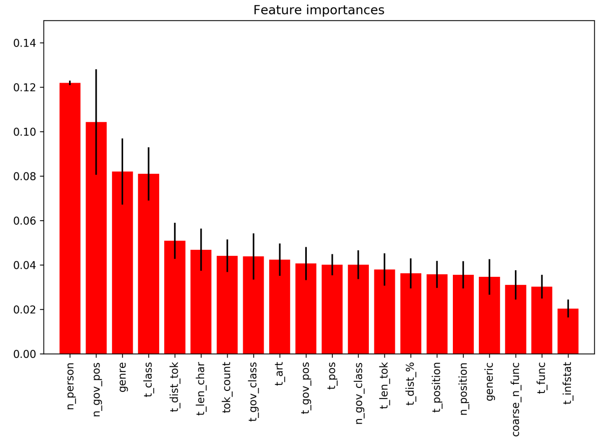

Figure 1: Variable importances for the classifier. Features beginning with n_ apply to the anaphor, and features with t_ to the antecedent.

To evaluate the importance of features in Figure 1 we use the Gini index of purity achieved at splits using each respective feature across all trees. Error bars indicate the standard deviation from the average importance across all trees in the ensemble. A Gini index of 0 means complete homogeneity (for our task, a 50-50 split on both sides), whereas 1 would mean perfect separation based on that feature. In addition to features discussed above, a feature 'generic' was introduced for phrases such as 'anyone', 'someone', 'somebody', etc. which behave differently from other PERSON entities, as well as a feature 't_art' coding the antecedent's article as definite, indefinite, demonstrative, or none.

The most important feature is 1st vs. 3rd person anaphor ('n_person'), as these are rather different situations: 1st person cases occur mainly with individuals speaking for aforementioned organizations, introduced as proper nouns (e.g. 'the SEC ... we' in (7)). Next is the POS tag of the anaphor's governor, which includes information about tense and can work in conjunction with verbs' semantic classes and grammatical functions (cf. Depraetere 2003, Annala 2008). Genre is surprisingly important in third place (cf. Levin 2001), indicating that settings licensing notional anaphora are genre specific. Replacing genre with a more coarse grained spoken/written variable degrades accuracy. Genre is closely followed by the semantic class of the antecedent, i.e. the entity in question, which is clearly relevant (+/-PERSON and more, see Section 4.2 for details).

Subsequent variables are less important, including distance, length and position in the document. Though both are helpful, using the article form ('t_art') is more important than the information status or previous mention ('t_infstat') based on antecedents to the antecedent (keeping in mind limitations of the coreference annotations, cf. Section 3.2). Grammatical functions are helpful, but less so than other features.

Looking at the actual classifications obtained by the classifier produces the confusion matrix in Table 2. The matrix makes it clear that the classifier is very good at avoiding errors against the majority class: it almost never guesses 'notional' when it shouldn't. Conversely, about 1/3 of actual notional cases are misclassified, predicted to be 'strict'. Among the erroneous cases, only 6 belong to Type III (about 15% of errors) , showing that the classifier largely handles this type quite well next to the other types, since Type III covers about 20% of plural-to-singular agreement cases.

Table 2: Confusion matrix for test data classification

		Predicted		
		Sg	Pl	Total
Actual	Sg	222	39	261
	Pl	7	81	88
	Total	229	120	349

4.2 Analysis of predictors

To understand why the features used in the previous section are helpful we analyze the distribution of notional anaphors for several non-obvious predictors individually. Beginning with processing factors, we can consider the effect of distance between anaphor and antecedent and position in the document, shown in Figures 2 and 3.

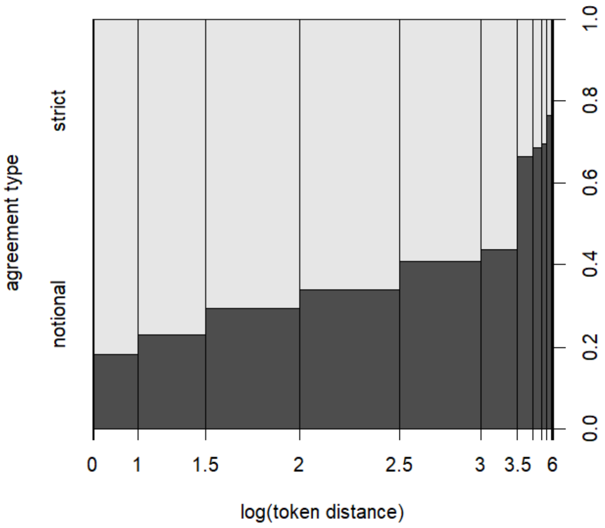

Figure 2: Log token distance between anaphor and antecedent.

In Figure 2, token distance is shown in log-scale, as greater distances are attested sparsely, and the breadth of each column in the spine plot corresponds to the amount of data it is based on. It is easy to see the perfectly monotonic rise in the proportion of notional agreement, beginning with under 20% at a log-distance of ˜1, all the way to over 50% at log-distances of ˜3.5 or higher (approximately 33 tokens and above).

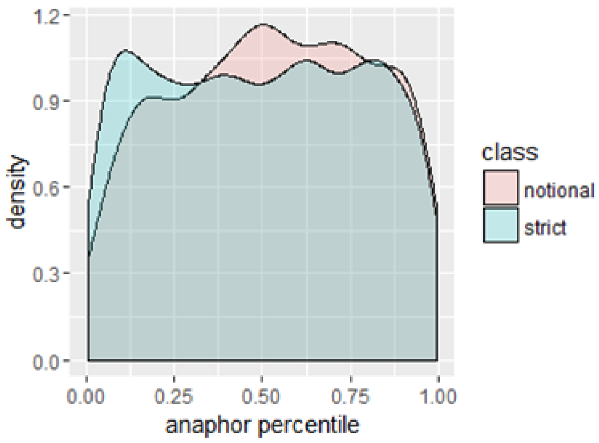

Figure 3: Position of anaphor as percentile of document length in tokens.

Figure 3 shows why position in the document matters: there is a slightly higher frequency of notional agreement after the halfway point of documents. This can be related to a speaker fatigue effect (speakers/writers become less constrained and exhibit less strict agreement as the document goes on), or due to editorial fatigue in written data (editors correct notional agreement, but notice it less frequently further in the document). However while we would only expect an editorial motivation to affect written data, the effect is found in both spoken and written documents, meaning a possible speaker fatigue effect cannot be discounted.

Next we can consider the effect of genre, and expectations that speech promotes notional agreement. This is confirmed in Table 3. However we note that individual genres do behave differently: data from the Web is closer to spoken language. The most restrictive genre in avoiding notional agreement is translations. Both of these facts may reflect a combination of modality, genre and editorial practice effects. However the strong differences suggest that genre is likely crucial to any model attempting to predict this phenomenon.

Table 3: Agreement patterns across genres

genre	**agreement**		
written	*notional*	*strict*	*% notional*
bible	169	487	25.76
newswire	344	843	28.98
translations	55	210	20.75
web	48	71	40.33
total written	616	1611	27.66
spoken	*notional*	*strict*	*% notional*
bc.conv	237	201	54.11
bc.news	296	378	43.91
phone	60	89	40.26
total spoken	593	668	47.02

Moving on to grammatical and semantic factors, we consider the entity type of the referring expression in Figure 4. The plot shows the chi square residuals for the association of each entity type with the two agreement types. Lines sloping top-right to bottom-left correspond to entity types preferring strict agreement (OBJECT, PLACE, PERSON), while top-left to bottom-right slopes correspond to types preferring notional agreement (QUANTITY, TIME, ORGANIZATION).

The result that PERSON somewhat prefers

strict agreement is surprising given the expectation that agentive, human-associated predicates have an effect promoting notional agreement (Depraetere, 2003). This is because many of those predicates were most often associated in our data with an ORGANIZATION telling, having or wanting to do something, and then being construed as a group of humans. This leads to the notional preference of the ORGANIZATION class. NPs actually classified as PERSON often included heads such as the very common *family* (mostly singular agreement), or potential Type III nouns which often take explicit gender (e.g. gender-specific *'baby ... her/his'*). Less surprising is the association of QUANTITY and TIME with notional agreement, covering cases such as *'a third of ... they'*, and counted time units in Type II phrases such as *'a couple of (minutes/hours)'*.

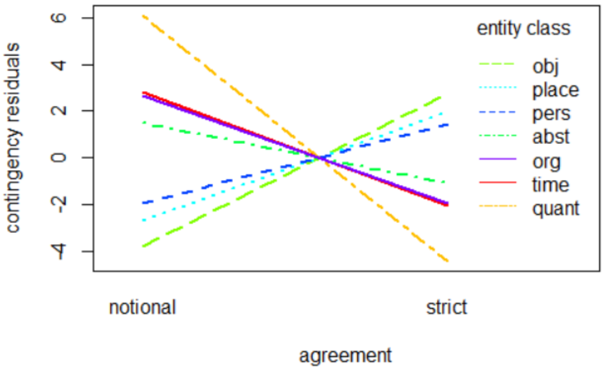

Figure 4: Chi square residuals for notional agreement by entity type. The legend is ordered by strictness.

Looking at the distribution of grammatical forms and functions, Table 4 shows imbalances based on the POS tag of the token governing the anaphor, and Figure 5 shows an association plot between dependency functions[2] and agreement patterns (rare POS and dependency labels have been omitted for clarity).

The table confirms the observations by Annala (2008) that present tense favors plural agreement more than past tense (VBD/VBN), but also reveals that nominal governors (NNP and more so NN, primarily possessed nouns of the entity in question), also promote singular agreement. This is echoed in the association plot in Figure 5. Possessive anaphors ('poss') prefer strict agreement and anaphoric subjects promote plural agreement, while the opposite is true for antecedents: if the antecedent is a subject, it is more likely to be realized later as a singular, and the opposite if it is a possessive.

Table 4: Agreement by anaphor governor POS

	notional	strict	% notional
VBG	112	94	54.36
VBpres[3]	218	255	46.08
VB	244	291	45.61
JJ	48	82	36.92
VBD	183	313	36.89
IN	65	117	35.71
VBN	81	163	33.19
NNP	8	18	30.76
NN	141	645	17.93

It is possible that the increased salience of subjects adds to speakers' tendency to refer back to them in keeping with the morphological number of the previous mention, while a late mention as a subject allows the salient anaphor position to select a disagreeing form more easily, without depending on previous mentions. Investigating this hypothesis further may require psycholinguistic data.

5 Conclusion

One of the fundamental challenges of notional agreement is the apparent unpredictability shown often in previous studies: the same nouns can appear under seemingly similar conditions with both types of agreement. The ensemble classifier presented here shows that despite this unpredictability, comparatively good predictions can be made on unseen data, with an accuracy of 86.81%, substantially improving on a baseline of 65.6%.[4]

[2] Two versions of the function labels were tested: coarse labels as used in Figure 5 (e.g. 'subj', 'clausal') and all available labels in the Stanford CoreNLP basic label set (distinguishing active 'nsubj' and passive 'nsubjpass', different types of clauses, etc.). The classifier works best with coarse labels for the anaphor's function but fine grained ones for the antecedent.

[3] The tags VBP and VBZ have been collapsed into VBpres, since they trivially imply whether the anaphor was singular or plural.

[4] An anonymous reviewer has asked to what extent state of the art coreference resolution systems also err on notional cases in general and the cases targeted here in particular: this is an interesting question which probably depends on the system, but it certainly seems possible that some architectures could benefit from notional agreement probability estimation, similarly to preprocessors predicting singleton status (Recasens et al., 2013) or other special constructions (e.g. anaphoric 'one' in English, Recasens et al. 2016).

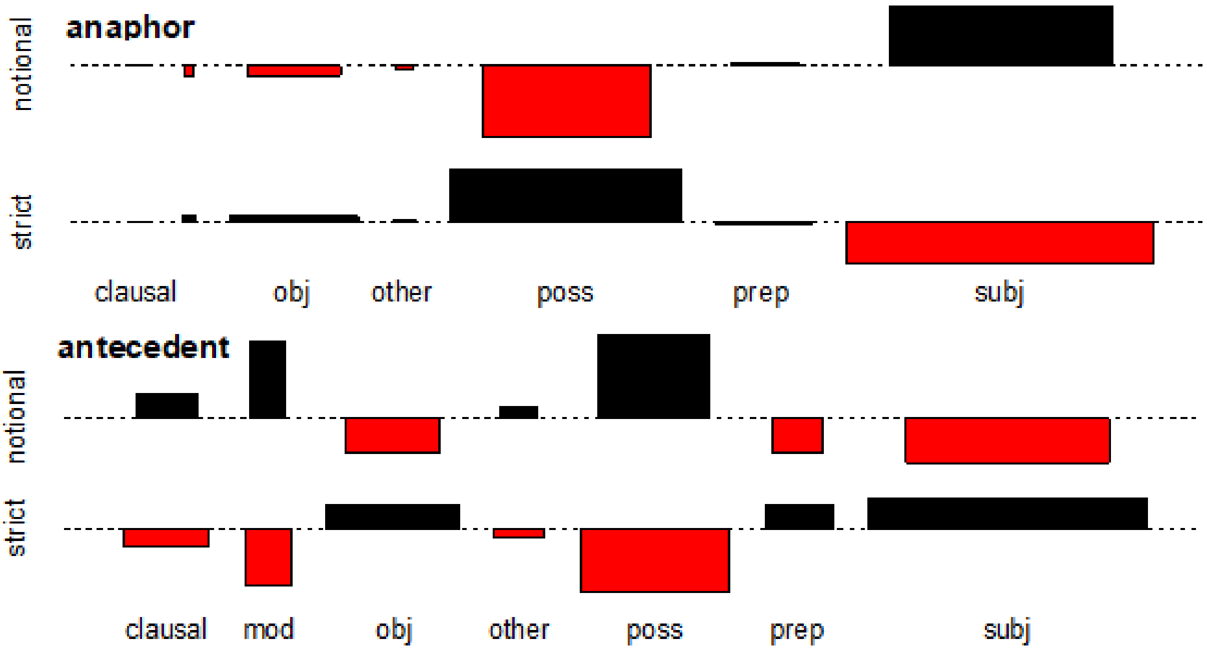

Figure 5: Association of pronoun choice and dependency functions of the anaphor and antecedent (top: anaphor; bottom: antecedent). The category ‘clausal’ collapses the labels ‘csubj’, ‘ccomp’ and ‘advcl’.

The classifier showed a good ability to recognize the majority class, but also learned to be ‘cautious’, guessing ‘strict’ in 1/3 of notional cases. A possible interpretation of this result is that for ambiguous cases, in which either form could be acceptable, the classifier chooses the safer majority class. In many such misclassified cases it seems likely that speakers would accept either variant, as in (16), which the classifier gets wrong:

(16) [Comsat Video, which distributes pay-per-view programs to hotel rooms], plans to add Nuggets games to [their] offerings

In this example, multiple signals suggest strict agreement, including an aforementioned, subject antecedent, and short distance to the 3rd person possessive anaphor. Based on features from the training data, it is a fair example of the environment of a ‘strict’ case; at the same time, it seems likely that speakers would accept a version with ‘its’, and it is not difficult to find similar examples, with similar distances, syntax and governing items, as in (17).[5]

(17) Ultimately, Lewis said, [her school] added African-American history to [its] offerings

Another aspect worth considering is the feature space used here, and some possible alternatives. Among the features tested but ultimately rejected in this study, we examined the presence of relative clause markers as suggested by Reid (1991), as well as some alternative semantic representations for governing verb semantics. For relative clauses, the importance of cases with ‘who’ as in example (4) turned out not to be useful in practice, despite the presence of well over 200 relative clauses in the data and over 150 with ‘who(m)’. It can be suspected that relative pronouns modifying the antecedent at the point it is mentioned have less interactions with anaphors, which can appear much later in the text, than with immediate subject-verb agreement cases which motivated the observation in Reid (1991).

For encoding verb semantics, the choice of VerbNet categories and the lack of disambiguation for ambiguous cases are both far from optimal. VerbNet classes do not necessarily map well onto verb groups’ preferences for notional agreement. It seems likely that other thematic, cluster-based or vector space-based methods of classifying verb semantics could be helpful for the present task. To this end we tested using semantic classes as assigned by the UCREL Semantic Analysis System (USAS, Rayson et al. 2004), which performed worse than VerbNet. Some VerbNet

[5] An anonymous reviewer has suggested that checking human acceptability of such deviating cases would be an interesting follow up study, and we certainly agree.

classes are mirrored in the USAS classes (e.g. communication verbs, the USAS coarse domain Q, or sub-classes in domain Q2); however in many cases it is possible that, by being much more specific (classes such as 'science and technology' in USAS), content domain classes encourage the classifier to memorize specific training instances, which do not generalize well. Ideally, a flexible semantic representation such as trainable embeddings would likely be helpful, but would require training on an external dataset beyond the notional agreement pairs, which only amount to a few thousand examples.

For future work, we can point out that while the classifier achieved overall good accuracy above chance, there is substantial room for improvement, and more features could be considered. These include phonological features (e.g. phonotactics around anaphors, metrical factors), morphological features (affixes, types of compounding), semantic features (more directly targeting predicates with distributive readings) and further context cues such as modifiers (adjectives, adverbs) and other words in the context not directly governing or governed by the noun in question. For NLP and NLG applications, it would be most useful to consider those variables for which we can build automatic taggers or generated contexts in real-time. At the same time, it will probably remain impossible to achieve perfect accuracy: it is expected that, as with many high level alternations, some element of inter- and even intra-speaker variation, as well as speakers' communicative intentions, will always create a certain degree of unpredictability in settings which are otherwise comparable.

References

Henri Annala. 2008. *Changes in Subject-Verb Agreement with Collective Nouns in British English from the 18th Century to the Present Day*. Pro gradu thesis, University of Tampere.

BBN Technologies. 2007. Co-reference guidelines for English OntoNotes. version 6.0. Technical report.

Anne Curzan. 2003. *Gender Shifts in the History of English*. Studies in English Language. Cambridge University Press, Cambridge.

Marcel den Dikken. 2001. "Pluringulars", pronouns and quirky agreement. *The Linguistic Review* 18:19–41.

Ilse Depraetere. 2003. On verbal concord with collective nouns in British English. *English Language and Linguistics* 7(1):85–127.

Kathleen Eberhard, J. Cooper Cutting, and Kathryn Bock. 2005. Making sense of syntax: Number agreement in sentence production. *Psychological Review* 112:531–559.

Morton Ann Gernsbacher. 1986. Comprehension of conceptual anaphora in discourse processing. In *Proceedings of the Eigth Annual Conference of the Cognitive Science Society*. Amherst, MA, pages 110–125.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63(1):3–42.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York, pages 57–60.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics* 38(1):173–218.

Geoffrey Leech and Jan Svartvik. 2002. *A Communicative Grammar of English*. Longman, London.

Magnus Levin. 2001. *Agreement with Collective Nouns in English*. Lund Studies in English 103. Lund University, Lund.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and Davide McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014: System Demonstrations*. Baltimore, MD, pages 55–60.

Ana E. Martinez-Insua and Ignacio M. Palacios-Martinez. 2003. A corpus-based approach to non-concord in present day English existential there-constructions. *English Studies* 84(3):262–283.

Laura L. Paterson. 2011. *The Use and Prescription of Epicene Pronouns: A Corpus-based Approach to Generic he and Singular they in British English*. Ph.D. thesis, Loughborough University.

Randolph Quirk, Sydney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the Workshop Beyond Named Entity Recognition: Semantic Labelling for NLP Tasks*. Lisbon, Portugal, pages 7–12.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of NAACL 2013*. Atlanta, GA, pages 627–633.

Marta Recasens, Zhichao Hu, and Olivia Rhinehart. 2016. Sense anaphoric pronouns: Am I one? In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL 2016*. San Diego, CA, pages 1–6.

Wallis Reid. 1991. *Verb and Noun Number in English: A Functional Explanation*. Longman, London.

Uli Sauerland. 2003. A new semantics for number. In *Proceedings of SALT 13*. CLC Publications, Ithaca, NY.

Nicholas Sobin. 1997. Agreement, default rules, and grammatical viruses. *Linguistic Inquiry* 28(2):318–343.

Adrian Staub. 2009. On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language* 60(2):1–39.

Matthew W. Wagers, Ellen F. Lau, and Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61:206–237.

Amir Zeldes and Shuo Zhang. 2016. When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proceedings of the NAACL2016 Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*. San Diego, CA, pages 92–101.