Journal of Artificial Intelligence Research 52 (2015) 445-475 Submitted 09/14; published 03/15

Modeling the Lifespan of Discourse Entities with Application to Coreference Resolution

Marie-Catherine de Marneffe MCDM@LING.OHIO-STATE.EDU
Linguistics Department
The Ohio State University
Columbus, OH 43210 USA

Marta Recasens RECASENS@GOOGLE.COM
Google Inc.
Mountain View, CA 94043 USA

Christopher Potts CGPOTTS@STANFORD.EDU
Linguistics Department
Stanford University
Stanford, CA 94035 USA

Abstract

A discourse typically involves numerous entities, but few are mentioned more than once. Distinguishing those that die out after just one mention (singleton) from those that lead longer lives (coreferent) would dramatically simplify the hypothesis space for coreference resolution models, leading to increased performance. To realize these gains, we build a classifier for predicting the singleton/coreferent distinction. The model's feature representations synthesize linguistic insights about the factors affecting discourse entity lifespans (especially negation, modality, and attitude predication) with existing results about the benefits of "surface" (part-of-speech and n-gram-based) features for coreference resolution. The model is effective in its own right, and the feature representations help to identify the anchor phrases in bridging anaphora as well. Furthermore, incorporating the model into two very different state-of-the-art coreference resolution systems, one rule-based and the other learning-based, yields significant performance improvements.

1. Introduction

Karttunen imagined a text interpreting system designed to keep track of "all the individuals, that is, events, objects, etc., mentioned in the text and, for each individual, record whatever is said about it" (Karttunen, 1976, p. 364). He used the term *discourse referent* to describe these abstract individuals. Some discourse referents are easily mapped to specific entities in the world, as with most proper names. Others are indeterminate in the sense that they are compatible with many different real-world entities, as with indefinites like *a train*. In either case, discourse referents can enter into anaphoric relations in discourse; even if we do not know exactly what real-world object *a train* picks out in *We heard a train in the distance* ..., we can nonetheless refer to it with subsequent pronouns and ascribe properties to it (... *It had a loud horn*).

Not all discourse referents enjoy repeat appearances in the discourse. Some lead long lives and appear in a wide variety of discourse contexts, whereas others never escape their birthplaces, dying out after just one mention. The central question of this paper is what factors influence the lifespan of a discourse referent. We focus on noun phrases, which are the most direct identifiers of discourse referents in English. More specifically, we seek to predict whether a given discourse referent will be *coreferent* (mentioned multiple times in a given discourse) or *singleton* (mentioned just once). The ability to make this distinction based on properties of the noun phrases used to identify these referents (henceforth, *mentions*) would benefit coreference resolution models, by simplifying the hypothesis space they consider when predicting anaphoric links, and it could improve performance on other tasks that require accurately tracking discourse entities, including textual entailment (Delmonte, Bristot, Piccolino Boniforti, & Tonelli, 2007; Giampiccolo, Magnini, Dagan, & Dolan, 2007) and discourse coherence (Hobbs, 1979; Grosz, Joshi, & Weinstein, 1995; Kehler, 2002; Barzilay & Lapata, 2008; Prasad, Dinesh, Lee, Miltsakaki, Robaldo, Joshi, & Webber, 2008).

The existing literature provides numerous generalizations relevant to the singleton/coreferent distinction. It is known, for example, that the internal syntax and morphology of the phrase used to establish the discourse referent provide important clues as to the lifespan of that referent (Prince, 1981a, 1981b; Wang, McCready, & Asher, 2006). Information structuring is also important; certain grammatical and discourse roles correlate with long lifespans (Chafe, 1976; Hobbs, 1979; Walker, Joshi, & Prince, 1997; Beaver, 2004). Features based on these insights have long been integrated into coreference resolution systems. Our contribution is to explore the interaction of all of these features with semantic operators like negation, modals, and attitude predicates (*know*, *be certain*, *wonder*). Such interactions were Karttunen's primary focus (Karttunen, 1973, 1976), and they have long dominated work on dynamic approaches to linguistic meaning (Kamp, 1981; Heim, 1982, 1992; Roberts, 1990; Groenendijk & Stokhof, 1991; Bittner, 2001). Here, we highlight the importance of such interactions for predicting the lifespans of discourse referents in actual text.

Our approach also capitalizes on the results of Durrett and Klein (2013) and Hall, Durrett, and Klein (2014) concerning the power of "surface" features for natural language processing (NLP) tasks. Those authors show that large sets of easily extracted part-of-speech (POS) and n-gram-based features can achieve results that are at least as good as those achieved with hand-engineered linguistic features. We therefore investigate the contribution of surface features for predicting the lifespan of discourse entities. We find that surface features alone have substantial predictive value for this task, but that adding more specialized linguistic features leads to reliable performance gains. This suggests that some of the linguistic constraints relevant for lifespan prediction go beyond what can be approximated with surface-level information given available data.

The first step in our analysis is to bring the insights from linguistic theories together into a single logistic regression model — the *lifespan model* — and assess their predictive power on real data. We show that the linguistic features generally behave as the existing literature leads us to expect, and that the model itself is effective at predicting whether a given mention is singleton or coreferent. The second step is to bring in surface features to obtain a more predictive model. We then provide an initial assessment of the engineering value of making the singleton/coreferent distinction by incorporating our lifespan model into two very different, state-of-the-art coreference resolution systems: the rule-based Stanford coreference system (Lee, Peirsman, Chang, Chambers, Surdeanu, & Jurafsky, 2011) and the learning-based Berkeley coreference system (Durrett & Klein, 2013). For both, adding our features results in a significant improvement in precision on the CoNLL-2011 and CoNLL-2012 Shared Task data, across all the standardly used coreference resolution measures, and we see reliable boosts in recall as well.

This article subsumes and extends the work of Recasens, de Marneffe, and Potts (2013). The specific differences are as follows. First, freed of NAACL's tight space constraints, we provide a much more in-depth linguistic analysis of the various features in our lifespan model, and include more details throughout. Second, we examine the contribution of surface features to the lifespan model. Third, we assess the value of the lifespan model for predicting which phrases will act as anchors in bridging anaphora. Fourth, to give a fuller evaluation of the coreference applications of our model, we incorporate our best lifespan model into a learning-based system (the Berkeley coreference system), complementing our previous results on the rule-based Stanford coreference system. Fifth, we use the most recent version of the CoNLL scorer (v8.0), which includes results according to BLANC and fixes a bug that incorrectly boosted B3 and CEAF scores by a few points. Sixth, we benefit from Kummerfeld and Klein's (2013) error analysis tool to gain deeper insights into the errors that our lifespan model helps with.

2. Linguistic Insights

This section briefly summarizes previous research on anaphora resolution, discourse structure, and discourse coherence in the linguistic literature. Our goal is to obtain a clear picture of how the lifespan of a discourse referent is shaped by features of its mentions — not only their local morphosyntactic features but also features of the syntactic and semantic environments in which they occur. The insights we gather in this section inform the design of the feature extraction functions in our lifespan model (Section 5) and in turn shape our contributions to the Stanford and Berkeley coreference systems (Section 8).

Karttunen (1976) was primarily concerned with the ways in which the semantic scope of an indefinite influences the lifespan of its associated discourse referent. In the three-sentence discourse (1), the indefinite *an exam question* in sentence 1 has text-level scope. As a result, its associated discourse referent is free to lead a long life, linking with a mention that is also at the text-level (sentence 2) and one that is embedded below negation (sentence 3).

(1) Kim read over *an exam question*. *It* was hard. He didn't understand *it*.

In contrast, as Karttunen observed, if an indefinite is interpreted in the scope of negation, then it is typically available for anaphoric reference inside that negative environment, as in (2), but not outside of it, as in (3). (We use # to mark discourses that are incoherent on the intended construal.)

(2) Kim didn't understand *an exam question* even after reading *it* twice.

(3) Kim didn't understand *an exam question*. #*It* was too hard.

Of course, (3) has a coherent construal on which *an exam question* is interpreted as taking wide-scope with respect to negation ('there is a question Kim didn't understand'). Such *inverse scope* readings are often disfavored, but they become more salient when modifiers like *certain* and *particular* are included (Fodor & Sag, 1982; Schwarzschild, 2002), or where the mention contains a *positive polarity item*, that is, an item like *some* or *tons of* that resists scoping under negation semantically (Baker, 1970; Israel, 1996, 2001):

(4) Kim didn't understand *a particular exam question*. She pondered *it* for hours to no avail.

(5) Kim didn't understand *some exam question*. She pondered *it* for hours to no avail.

Conversely, using a *negative polarity item* (NPI) like *any* inside the indefinite mention essentially ensures a narrow-scope reading (Ladusaw, 1996; Israel, 2004), which leads to an impossible-to-resolve anaphoric link for simple variants of (3):

(6) Kim didn't understand *any exam question*. #*It* was too hard.

The pattern Karttunen saw in all this is that semantic scope and anaphoric potential are intimately related: a given mention can participate in anaphoric relationships within its scope, but not outside of it. Broadly speaking, this is familiar from quantificational binding in logical languages (Cresswell, 2002) and variable scope in the control structures of programming languages (Muskens, van Benthem, & Visser, 1997). Thus, an indefinite with text-level scope has free reign, whereas one inside the scope of an operator like negation is restricted to links that do not span the outer boundaries of that scopal environment. These are semantic generalizations that might not be directly reflected in the surface syntax, but interpretive preferences and internal morphosyntactic features of the mention can help to disambiguate the intended logical form.

Karttunen (1976) immediately generalized his observations about negation and discourse reference to modal auxiliaries and non-factive attitude predicates like *want* and *claim*. The following are based on his original examples:

(7) Bill can make *a kite*. #*It* has a long string.

(8) John wants to catch *a fish*. #Do you see *it* from here?

(9) Sandy claims that Jesse bought *a bicycle*. #*It* has a green frame.

As with negation, the pattern makes intuitive sense. Bill's abilities regarding kite construction do not involve any specific kite, and hence the first sentence of (7) does not automatically establish the right sort of discourse referent. Similarly, wanting to catch a fish does not guarantee the salience (or even existence) of a fish, and Sandy might be so unreliable as a source that *a bicycle* has no status outside of the semantic scope of *claim*.

All of (7)–(9) cohere if the indefinite is interpreted outside of the scope of the relevant semantic operator. The relative preferences for surface and inverse scope are harder to characterize than they were with negation, because they are influenced in complex ways by the semantics and pragmatics of the attitude predicate, the reliability of the source of the information, and the nature of the conversational issues and goals. For example, if the speaker of (9) regards Sandy as a reliable source regarding Jesse's bike buying, then *a bicycle* will likely attain text-level scope as a by-product of 'Jesse bought a bicycle' becoming a text-level commitment. Karttunen (1973) discusses these patterns, observing that, in many contexts, pragmatic pressures encourage embedded content to become elevated to the text level in this way. De Marneffe, Manning, and Potts (2012) study newspaper data in which this is an extremely common pattern because the attitude verbs tend to function as evidential markers for the source of the embedded content (Rooryck, 2001; Simons, 2007). We will see later that attitude predicates seem to encourage long lifespans in the OntoNotes data too (the majority of which is news-like), arguably as a result of just these pragmatic factors.

We have so far restricted attention to anaphoric links in which an indefinite establishes a new discourse referent and a pronoun refers to it. Our observations carry over directly to links from indefinites to definite noun phrases, which linguistic theories treat roughly as pronouns with additional descriptive content (for discussion, see the work of Elbourne, 2008). Other mention-patterns tend to be quite different, though. Where discourse referents are established by definites or named entities, the interactions with negation and other operators are simpler because definites and named entities do not interact scopally with these operators (but see the work of Aloni, 2000, for related issues involving presupposition and intensionality). Thus, such anaphoric connections are unconstrained by the factors we have been discussing. Conversely, truly quantified phrases like *no student* and *every linguist* are severely limited, not only by their interaction with other operators but also by their own deficiencies when it comes to establishing discourse referents. There are cases in which these expressions establish new discourse referents, but they seem to be infrequent and unusual (Wang et al., 2006).

Cross-cutting the above considerations are factors that have long been central to studies of coreference and anaphora within computational linguistics and NLP. For instance, animate nouns are generally the most likely to lead long discourse lives, whereas mentions that refer to abstract objects like quantities, percentages, and other measures tend to be singleton. We assume that these statistical patterns derive, not from narrow linguistic constraints, but rather from general cognitive biases concerning how people conceptualize and discuss different kinds of objects. However, there is evidence that these biases can make their way into the grammars of specific languages in the form of morpho-semantic phenomena like obviation (Aissen, 1997) and differential object marking (Aissen, 2003).

The syntactic environment in which the phrases occur will also modulate their anaphoric potential and hence their lifespans. For example, Prince (1981b) reports that semantically indefinite phrases using *this*, as in *There was this guy in the back row*, are highly likely to be referred to in a subsequent clause. Similarly, Chafe (1976) shows that information structuring choices are also predictive of whether a given noun phrase will serve as the antecedent for later referential devices. There are also close correlations between being in a syntactic topic position and leading a long discourse life (Grosz et al., 1995; Beaver, 2004); for a focused evaluation of these ideas for handling coreference, see the work of Beaver (2007).

We seek to incorporate all of the above observations into our lifespan model. There are additional patterns from the literature that we do not pursue, because they are too infrequent in our data. For example, Karttunen (1976) also identified a natural class of counterexamples to his basic scope generalizations: certain sequences of intensional predicates support exceptional anaphoric links, a phenomenon that was later studied systematically under the heading of *modal subordination* (Roberts, 1990, 1996):

(10) Frank wants to marry *a rich linguist*. #*She* is kind.

(11) Frank wants to marry *a rich linguist*. *She* should be kind.

In addition, mentions inside parenthetical clauses are less likely to introduce long-term discourse referents, due to the likelihood that the parenthetical clause itself conveys only secondary content as compared with the main clause that hosts it (Potts, 2005). Thus, while anaphoric links into and out of parentheticals are possible (AnderBois, Brasoveanu, & Henderson, 2010; Potts, 2012), they seem to arise relatively rarely, a valuable piece of practical advice for appositive-rich texts like scientific papers but unfortunately not one we could put into action here.

Karttunen's observations helped set the agenda for dynamic approaches to semantics for the next few decades (Kamp, 1981; Heim, 1982; Groenendijk & Stokhof, 1991). That literature refined and extended his observations in numerous ways. Taken together, the findings suggest that intensional operators and negation interact in complex ways with discourse anaphora. By default, we expect phrases introduced in the scope of such operators to lead short lifespans, but it is possible for them to take wide-scope with respect to those operators, which broadens the range of anaphoric links they can establish. Such readings are favored or disfavored by the pragmatics of the situation as well as the lexical and syntactic nature of the phrases involved. In what follows, we seek to model these interactions and use them to inform a lifespan model.

3. Previous Engineering Efforts and Quantitative Evaluations

The above insights have inspired NLP researchers to try to predict the roles that different mentions will play in coreference chains. Previous work in this area can be subdivided into detecting four different targets: non-referential mentions, non-anaphoric mentions, discourse-new mentions, and non-antecedent mentions. The terminology has not always been used in a consistent way in linguistics or NLP, but we believe that the results can ultimately be brought together. Here, we aim to clarify the terminology and find common insights behind the various features that have been used. We are the first to single out the singleton/coreferent detection task as such, but our work finds important antecedents in the existing literature.

3.1 Non-referential Mentions

Some noun phrases do not refer to a discourse referent but rather just fill a syntactic position. In English, the canonical example of a non-referential NP is the expletive pronoun *it*, as in *It is obvious that we will succeed*. Some lexical NPs do not introduce a discourse referent either, such as *a linguist* in *Pat is a linguist*: while the mention *Pat* does introduce a discourse referent, *a linguist* simply predicates something of her. Detecting such non-referential uses plays a role in coreference resolution: since these NPs do not pick out discourse referents (new or existing), they cannot enter into any anaphoric relations of the kind under consideration here.

Early work in non-referentiality detection focuses on the pronoun *it*, aiming to distinguish referential uses from non-referential ones. Paice and Husk (1987) develop a rule-based system, Evans (2001) uses a supervised approach, and Müller (2006) focuses on the use of *it* in spoken dialog. These studies mainly employ lexico-syntactic features of the immediate surrounding context of the pronoun. Similarly, Bergsma, Lin, and Goebel (2008) explore a system that uses Web-count features derived from the Google n-grams data (Brants & Franz, 2006) to capture the most frequent subjects that can replace the pronoun *it*: for referential cases (e.g., *it is able to*), other words than *it* will be frequent in the n-grams, such as ***he** is able to* or ***China** is able to*, whereas for non-referential cases, the pronoun *it* will likely be the most frequent subject (e.g., ***it** is important to*).

More recently, Bergsma and Yarowsky (2011) develop the NADA system, which improves on Bergsma et al. (2008) by incorporating lexical features. The lexical features indicate the presence or absence of some strings at specific positions around the pronoun: three-grams to five-grams spanning the pronoun; two tokens before the pronoun to five tokens after the pronoun with their positions; any token within twenty tokens to the right of the pronoun; and any token within ten tokens to the left of the pronoun that is a named entity or belongs to the following list: *that*, *this*, *and*, *said*, *says*, *it*, *It*, *its*, *itself*. Using both types of features, lexical and Web-count, they achieve 85% accuracy on different datasets.

Byron and Gegg-Harrison (2004) apply some of the linguistic insights highlighted by Karttunen (Section 2) to the special case of pronoun resolution, seeking to discard non-referential indefinite NPs from the set of potential antecedents for pronouns. They use a hard filter for non-referential mentions, looking at the presence of indefinites, negation, apposition (hand-labeled), modals, adjectival phrases or predication adjuncts (tagged '-CLR' in the Penn Treebank), predicates of copular verbs (tagged '-PRD'), and noun phrases that express a value. They found that removing non-referential mentions gave a small boost in performance for pronoun resolution.

3.2 Non-anaphoric Mentions

Non-anaphoric NPs are those whose interpretation does not depend on a previous mention in the text. For example, the phrase *the new Scorsese movie that stars De Niro* in (12) (while manifesting many kinds of context dependence) does not depend on any other overt phrases in order to capture all of its descriptive content. In contrast, *the movie* in (13) crucially links back to the previous sentence for its descriptive content; it superficially involves just the predicate 'movie', but it is construed as having the additional property 'seen by the speaker the previous night'.

(12) Last night, I watched *the new Scorsese movie that stars De Niro*.

(13) Last night, I watched a movie and read a paper. *The movie* was directed by Scorsese.

There is no direct correspondence between anaphora and coreferentiality. Coreferent mentions can be non-anaphoric (as in a text containing multiple tokens of the phrase *The White House*), and anaphoric mentions can be coreferent or non-coreferent (van Deemter & Kibble, 2000). Cases of *bridging anaphora* (Clark, 1975) like (14) involve non-coreferent anaphora. Here, *the ceiling* is interpreted as the ceiling of the room mentioned in the previous sentence, and thus it is anaphoric to *the room* without being coreferent with it or any other phrase in the discourse.

(14) I looked into **the room**. *The ceiling* was very high.

We return to such cases in Section 6, where we use our lifespan model to characterize the sense in which bridging anchors like *the room* lead longer lifespans than a count of their strictly coreferent mentions would suggest.

Poesio, Uryupina, Vieira, Alexandrov-Kabadjov, and Goulart (2004) and Poesio, Alexandrov-Kabadjov, Vieira, Goulart, and Uryupina (2005) summarize previous approaches to non-anaphoricity detection, which they refer to as *discourse-new* detectors. Vieira and Poesio (2000) focus on definite NPs and use syntactic heuristics based on pre- and post-modification to distinguish between anaphoric and non-anaphoric NPs. Modification is a good indicator of anaphoricity; heavily modified phrases like *the new Scorsese movie that stars De Niro* tend to be non-anaphoric, whereas short phrases with general descriptive content like *the movie* tend to be anaphoric. Bean and Riloff (1999) also focus on definite NPs: in addition to syntactic heuristics based on pre- and post-modification, they use techniques mining lists of likely non-anaphoric NPs (such as the presence of NPs in the first sentence of a document). Compared to Vieira and Poesio (2000), they obtain substantially higher recall (with recall and precision figures around 80%).

In their non-anaphoricity detector, Poesio et al. (2005) use a head feature (distance between NPs with identical heads), syntactic features (e.g., occurring inside an appositive or copular clause, being post-modified), capitalization of the mention, presence of the mention in the first sentence of a Web page, position of the mention in the text, and the probability of the mention being definite as computed from the Web using the technique of Uryupina (2003). They find that the most important features are the head feature and the definiteness probabilities.

3.3 Discourse-New Mentions

Discourse-new mentions are those that introduce a new entity into the discourse (Prince, 1981b; Fraurud, 1990). The entity might be singleton or involve a chain of coreferring mentions in which the first phrase is the discourse-new one and the rest are considered discourse-old. Cast as an information status task, the goal of discourse-new mention detection is to find discourse referents which were not previously available to the hearer/reader; e.g., see the work of Nissim (2006).

Ng and Cardie (2002) develop a discourse-new classifier that targets every kind of NP using a variety of feature types: lexical (string and head matching, conjunction), morpho-syntactic (definiteness, quantification, number), grammatical (appositional or copular context, modifier structure, proper-noun embedding), and shallow semantic (e.g., WordNet features). They incorporate the classifier into their coreference resolution system, pre-filtering NPs that are tagged as discourse-new. However, this pre-filtering ultimately hurts coreference resolution system performance: even though precision increases, recall drops considerably. In Section 8.2.3, we report similar results for our model instantiated with discourse-new pre-filtering, but we find that the recall drop can be avoided if filtering is applied only when the mention under analysis is tagged as discourse-new and the antecedent candidate is tagged as singleton.

Ng and Cardie's (2002) work is cast as *non-anaphoricity* detection, but their model is perhaps better described as trying to distinguish coreferent mentions from those that are singleton or initiate coreference chains. More specifically, they write, "a positive instance is created for each NP that is involved in a coreference chain but is not the head of the chain" (Ng & Cardie, 2002, p. 3), which picks out non-initial members of coreference chains. Conversely, "a negative instance is created for each of the remaining NPs" (Ng & Cardie, 2002, p. 3), i.e., those without any antecedents.

Uryupina (2009) proposes a discourse-new mention detector for any kind of NP. The classifier relies on features falling into three categories she defines: 'lexical' (number of words in the mention), 'syntactic' (POS, number, person, determiner, pre- and post-modification), 'semantic' (gender, semantic class), and 'salience' (grammatical role, position in the sentence and in the paragraph). In addition, she includes some of Karttunen's features as implemented by Byron and Gegg-Harrison (2004). Her classifier also checks for mentions with identical heads, and distance between these. Only the syntactic and head features deliver improvements to a majority baseline (which marks each NP as discourse-new), performing almost as well as all the features together. Uryupina notes, however, that most of the features, and especially those based in Karttunen's ideas, have not been designed for discourse-new mention detection.

Both Ng and Cardie (2002) and Uryupina (2009) integrated their discourse-new detector into a coreference resolution system in a pipeline manner. For a joint approach to discourse-new detection and coreference resolution, see the work of Denis and Baldridge (2007).

3.4 Non-antecedent Mentions

As Uryupina (2009) observes, for coreference resolution, what matters is the fact that some NPs are unavailable as antecedents. She therefore builds a classifier that marks NPs as likely antecedents or not. Her system is based on the same features as her discourse-new detector described above (Section 3.3). For non-antecedenthood detection, only the syntactic and semantic features lead to a significant precision improvement over a majority baseline (which marks each NP as non-antecedent), with the syntactic features alone performing as well as all the features together.

3.5 Our Approach: Singletons

Our model cross-cuts these four categories. Unlike previous models of non-referentality, ours is not restricted to pronouns or to indefinite NPs, but tries to identify any kind of non-referential NP as well as any referential NP whose referent is mentioned only once (i.e., singleton). Thus, all

			MENTIONS	
Dataset	Docs	Tokens	Coreferent	Singletons
Training	2,802	1.3M	152,974	181,274
Development	343	160K	18,855	23,140
Test	348	170K	19,407	23,657

Table 1: CoNLL-2012 Shared Task data statistics. We added singletons (noun phrases not annotated as coreferent), which account for 55% of the referents in the development set.

non-referential NPs fall into our singleton class. On the other hand, there is no strict correspondence between our singleton/coreferent distinction and the non-anaphoric/anaphoric distinction, since anaphoricity is based on whether the mention relies on a previous one for its interpretation, whereas the singleton/coreferent divide is based on how long the lifespan of an entity is. Similarly, discourse-new mentions can either be coreferent or singleton in our classification, depending on whether the entity is mentioned again or not.

In terms of feature representations, we have tried to stay as close as possible to Karttunen's original insights: we extract the features from full syntactic parses, seeking to remain faithful to the underlying semantic relationships involved, and we include feature interaction terms to capture the complex set of dependencies reviewed above in Section 2. This approach allows us to both evaluate those linguistic ideas quantitatively and to assess their practical contributions to full coreference systems.

4. Data

The data used throughout this paper come from the CoNLL-2012 Shared Task data (Pradhan, Moschitti, Xue, Uryupina, & Zhang, 2012), which included the 1.6M English words from OntoNotes v5.0 (Pradhan & Xue, 2009) with several common layers of annotation (coreference, parse trees, named-entity tags, etc.). The OntoNotes corpus contains documents from seven different domains: broadcast conversation (20%), broadcast news (13%), magazine (7%), newswire (21%), telephone conversation (13%), weblogs and newsgroups (15%), and pivot text (11%). Most of these genres are news-like, with the exception of the pivot texts (which come from the New Testament) and the telephone conversations. We used the training, development, and test splits as defined in the shared task (Table 1). Since the coreference annotations of OntoNotes do not contain any singleton mentions, we automatically marked as singleton all the noun phrases not annotated as coreferent. We excluded verbal mentions.

Because we mark as singleton all the noun phrases not annotated as coreferent, our definition of singletons includes non-referential noun phrases such as *it* in *It is raining*, and *president* in *He served as president for two terms* (Section 3.1). This makes practical sense: the starting point of most coreference resolution systems is to take all noun phrases as possible candidates for coreference and subsequently find the clusters that are coreferent with one another. The more phrases we can accurately identify as singleton, the more phrases we can exclude from this clustering step, which should translate directly into performance gains.

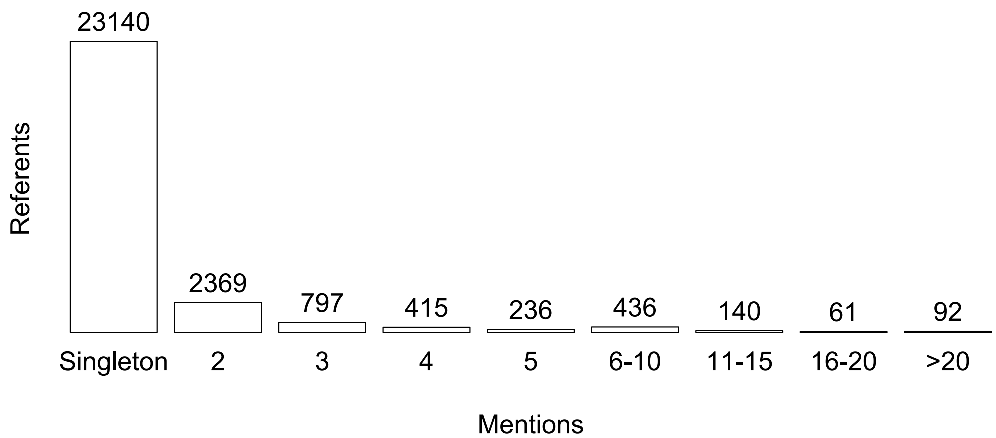

Figure 1: Distribution of referent lifespans in the 2012 OntoNotes development set.

5. Predicting Lifespans with Linguistic Features

We now describe our model for predicting the lifespan of discourse referents using the linguistic factors proposed in Section 2. The model makes a binary distinction between discourse referents that are not part of a coreference chain (singleton) and those that are part of one (coreferent). The distribution of lifespans in our data is shown in Figure 1.

This plot gives the number of entities associated with a single mention, the number associated with two mentions, and so forth. The fact that singletons so dominate the data suggests that the binary singleton/coreferent division is a natural one. The propensity toward singletons also highlights the relevance of detecting singletons for a coreference system. Following Bergsma and Yarowsky (2011), we use a logistic regression model, which has been shown to perform well on a range of NLP tasks. We fit the logistic regression model in R (R Development Core Team, 2013) on the training data, coding singletons as '0' and coreferent mentions as '1'. Thus, throughout the following tables of coefficient estimates, positive values favor coreferent mentions and negative values favor singletons. We turn now to describing and motivating the features of this model.

5.1 Morphosyntax of the Mention

Table 2 summarizes the features from our model that concern the internal morphology and syntactic structure of the mention, giving their coefficient estimates. In all the tables, if not indicated otherwise, the coefficient estimates are significant at $p < 0.001$. We use $*$ to indicate significance at $p < 0.05$, and † to indicate estimates with $p \geq 0.05$. The morphosyntactic features include type ('pronoun', 'proper noun', 'common noun'), animacy, named-entity tag, person, number, quantification ('definite', 'indefinite', 'quantified'), and number of modifiers of the mention. Many are common in coreference systems (Recasens & Hovy, 2009), but our model highlights their influence on lifespans. Where available, we used gold annotations to derive our features, since our primary goal is to shed light on the relevance of the features claimed to influence lifespans.

The morphosyntactic features were operationalized using static lists and lexicons as well as the Stanford dependencies as output by the Stanford parser (version 2.0.3; de Marneffe, MacCartney, & Manning, 2006) on the gold constituent trees. The features are extracted in the following way:

Type The type feature captures whether the mention is a pronoun, proper noun, or common noun. The value is determined by the gold POS tag of the mention and its named-entity tag.

Animacy We set animacy values ('animate', 'inanimate', 'unknown') using a static list for pronouns, named-entity tags (e.g., PERSON is animate whereas LOCATION is not), and a dictionary bootstrapped from the Web (Ji & Lin, 2009).

Person Person values ('1', '2', '3') are assigned only to pronouns (identified by POS tag), using a static list. Mentions that are not pronouns get a value of '0'.

Number The number value ('singular', 'plural', 'unknown') is based on a static list for pronouns, POS tags, Bergsma and Lin's (2006) static dictionary, and named-entity tags. (Mentions marked as a named entity are considered singular with the exception of organizations, which can be both singular and plural and get the value 'unknown'.)

Quantification As we discussed in Section 2, indefinites and definites can be given a referential semantics that pairs naturally with discourse anaphora, whereas the anaphoric possibilities of truly quantified terms are restricted. To operationalize quantification and decide whether a mention is definite, indefinite, or quantified, we use the dependencies to find possible determiners, possessors, and numerical quantifiers of a mention. A mention is 'definite' if it is a named entity, if it has a possessor (e.g., *car* in *John's car* is definite), or if its determiner is definite (*the*), demonstrative, or possessive. A mention is 'quantified' if it has a numerical quantifier (e.g., *two cars*) or if its determiner is *all*, *both*, *neither* or *either*. All other mentions are 'indefinite'.

Number of modifiers We added a feature counting how many modifiers the mention has, seeking to capture a correlation with specificity and referentiality. As modifiers, we counted adjectival, participial, infinitival, and prepositional modifiers as well as relative clause modifiers, noun compounds, and possessives. (Thus, there are four modifiers in the phrase *a modern multifunctional business center costing 60 million yuan*.)

Named entities Our model also includes named-entity features for all of the 18 OntoNotes entity-types, with 'NER = O' true of non-named-entities. We used the gold entity-type annotation.

Table 2 summarizes the coefficient estimates we obtain for these features. In broad terms, the picture is as one would expect from the taxonomy of given and new defined by Prince (1981b) and assumed throughout dynamic semantics (Kamp, 1981; Heim, 1982): pronouns depend on anaphoric connections to previous mentions for disambiguation and thus are likely to be coreferent. This is corroborated by the positive coefficient estimate for 'Type = pronoun'.

Few quantified phrases participate in discourse anaphora (Partee, 1987; Wang et al., 2006), accounting for the association between quantifiers and singletons (as measured by the negative coefficient estimate for 'Quantifier = quantified').

The negative coefficient for indefinites is initially surprising. As seen in Section 2, theories stretching back to Karttunen (1976) say that indefinites excel at establishing new discourse entities and so should be frequent participants in coreference chains, but here the association with such

Feature	Coefficient		Feature	Coefficient	
Type = pronoun	1.17		NER = GPE	3.46	
Type = proper noun	1.89		NER = LANGUAGE	2.56	
Animacy = inanimate	−1.36		NER = LAW	2.85	
Animacy = unknown	−0.39		NER = LOCATION	2.83	
Person = 1	1.04		NER = MONEY	0.05	†
Person = 2	0.13		NER = NORP	0.82	
Person = 3	1.62		NER = O	4.17	
Number = singular	0.61		NER = ORDINAL	−0.90	
Number = unknown	0.17		NER = ORGANIZATION	3.39	
Quantifier = indefinite	−1.43		NER = PERCENT	0.88	
Quantifier = quantified	−1.25		NER = PERSON	2.28	
Number of modifiers	−0.39		NER = PRODUCT	2.64	
NER = DATE	1.83		NER = QUANTITY	−0.02	†
NER = EVENT	2.89		NER = TIME	1.53	
NER = FACILITY	2.94		NER = WORK OF ART	2.42	

Table 2: Internal morphosyntactic features of the lifespan model. † indicates a non-significant coefficient ($p \geq 0.05$); no sign indicates a significant coefficient ($p < 0.001$).

chains is negative. We return to this in Section 5.3, where we argue that interactions with semantic operators explain this fact.

The behavior of the named-entity (NER) features is closely aligned with previous models and our theoretical discussion above. As a rule, named entities behave like 'Type = proper noun' in associating with coreferent mentions. The exceptions are MONEY, ORDINAL, NORP (for nationalities and religions), PERCENT, and QUANTITY, which seem intuitively unlikely to participate in coreference chains. The person, number, and animacy features together suggest that singular animates are excellent coreferent noun phrases.

The one real surprise for us here concerns the feature 'Number of modifiers'. Inspired by observations of Fodor and Sag (1982) and Schwarzschild (2002), we expected this feature to positively correlate with being coreferent. Our reasoning was that increased modification would likely result in increased specificity, thereby making the associated discourse referent more identifiable and more distinctive. The opposite seems to hold in our data. However, we hesitate to conclude from this that the original hypothesis is mistaken. Rather, we suspect that our model is just insufficiently sensitive to interactions between modifier counts and the lexical semantics of the modifiers themselves.

5.2 Grammatical Role of the Mention

Synthesizing much work in Centering Theory and information structuring, we hypothesized that coreferent mentions are likely to appear as core verbal arguments and favor sentence-initial (topic-tracking) positions (Ward & Birner, 2004). To capture these insights, we used the grammatical relation of the mention given by the Stanford dependencies on gold constituents, and the sentence position of the mention.

Feature	Coefficient		Feature	Coefficient
Sentence Position = end	−0.22		Relation = noun argument	0.56
Sentence Position = first	0.03	†	Relation = other	−0.67
Sentence Position = last	−0.31		Relation = root	−0.61
Sentence Position = middle	−0.11		Relation = subject	0.65
In coordination	−0.48		Relation = verb argument	0.32

Table 3: Grammatical role features of the lifespan model. † indicates a non-significant coefficient ($p \geq 0.05$); no sign indicates a significant coefficient ($p < 0.001$).

Sentence position Sentence position was determined based on the raw string: 'first' indicates that the mention is the first word of the sentence, 'end' the last word, and 'begin', 'middle', and 'last' indicate whether the mention is situated in the first, second, or last third of the sentence, respectively.

Relation To distinguish among grammatical relations, we check whether the mention is a 'subject', 'adjunct' (which includes prepositional objects, adverbial modifiers, and temporal modifiers), 'verb argument' (which includes direct and indirect objects, clausal complements, adjectival complements and attributes), or 'noun argument' (which includes relative clauses, appositions, possessives, noun compounds, and adjectival modifiers).

In coordination We also indicated whether or not the mention is a conjunct to see whether being inside a coordinate phrase affects coreference in ways that go beyond the grammatical role of the containing phrase.

The coefficient estimates in Table 3 support our general hypotheses: arguments make good discourse referents, subjects best of all, whereas sentence-final positions disfavor coreference. In addition, we note that the model identifies a negative correlation between coordination and coreference.

5.3 Semantic Environment of the Mention

Table 4 highlights the complex interactions between discourse anaphora and semantic operators introduced in Section 2. These interactions have been a focus of logical semantics since Karttunen (1976), whose guiding observation is semantic: an indefinite interpreted inside the scope of a negation, modal, or attitude predicate is generally unavailable for anaphoric reference outside of the scope of that operator. Heim (1992) also relates the anaphoric properties of NPs to scope-taking and the entailments of attitude predications.

We do not have direct access to semantic scope, but we expect syntactic scope to correlate strongly with semantic scope. We therefore used dependency representations to define features capturing syntactic scope for negation, modal auxiliaries, and a broad range of attitude predicates (181 verbs and 374 nouns from Saurí, 2008). Technically, for a given mention, we produce a 'negation', 'modal' or 'under attitude verb' feature according to the presence of pre-defined negation or modality markers (such as *not*, *can*, *may*) or attitude predicates (e.g., *accuse*, *allege*, *doubt*, *say*) in the dependency path. For example, the NP *relief* will be given a 'negation' feature in *while the financial storm shows no sign of relief today*, since it is under the scope of *no sign*. Similarly, the mention *scientific and technological companies* is in the scope of the modal auxiliary *would* and the

Feature	Coefficient	
Presence of negation	−0.18	
Presence of modality	−0.22	
Under an attitude verb	0.10	
AttitudeVerb * (Type = pronoun)	0.41	
AttitudeVerb * (Type = proper noun)	0.10	
AttitudeVerb * (Quantifier = indefinite)	−0.19	
AttitudeVerb * (Quantifier = quantified)	0.10	†
Modal * (Type = pronoun)	0.13	∗
Modal * (Type = proper noun)	0.35	
Modal * (Quantifier = indefinite)	−0.00	†
Modal * (Quantifier = quantified)	0.17	†
Negation * (Type = pronoun)	1.07	
Negation * (Type = proper noun)	0.30	
Negation * (Quantifier = indefinite)	−0.36	
Negation * (Quantifier = quantified)	−0.39	†
Negation * (Number of modifiers)	0.11	

Table 4: Semantic environment features and interactions in the lifespan model. † indicates a non-significant coefficient ($p \geq 0.05$); no sign indicates a significant coefficient ($p < 0.001$); $*$ indicates significance at $p < 0.05$.

attitude verb *said* in *firms from Taiwan said that they would establish scientific and technological companies in the zone*, and so it receives 'modal' and 'under attitude verb' features.

Table 4 summarizes our model's semantic environment features and their interactions. The interaction terms added to the model follow the previous linguistic literature: we expect that the scope of the semantic operators (negation, modality and attitude predicate) will interact with the internal syntax of the mention, specifically with its type and its definiteness/quantification. The results are beautifully aligned with our guiding linguistic hypotheses. First, negation and modality both negatively correlate with coreference, as expected given the constraints they impose on lifespans. Interacting these semantic features with those for the internal syntax of mentions also yields the expected results: since proper names and pronouns are not scope-taking, they are largely unaffected by the environment features, whereas indefinites, which are affected by scope, emerge as even more restricted, just as Karttunen and others would predict.

The coefficient values for attitude predicates and their interactions seem anomalous in light of the semantics of these items. In Section 2, we noted that non-factive attitude predicates like *say* cannot offer semantic guarantees that mentions in their scope will survive outside that scope. This might lead one to think that they will be biased against long-lived mentions, when in fact we see the opposite. However, we also observed that pragmatic factors often facilitate exceptional anaphoric dependencies in attitude predications. Karttunen (1973) referred to this as the 'leakiness' of these predicates — information introduced in their scope seems often to percolate up to the text level in a wide range of contexts (Rooryck, 2001; Simons, 2007; Harris & Potts, 2009). Since the lifespan

	# FEATURES	SINGLETON			COREFERENT			ACCURACY
		Recall	Precision	F1	Recall	Precision	F1	
LINGUISTIC	123	80.2	77.5	78.8	71.4	74.6	73.0	76.3
SURFACE	73,393	80.2	79.9	80.0	75.3	75.6	75.4	78.0
COMBINED	73,516	81.1	80.8	80.9	76.4	76.6	76.5	79.0
CONFIDENT	73,516	56.0	89.8	69.0	48.2	90.7	62.9	52.5

Table 5: Recall, precision, F1 and accuracy for the three different sets of features on the OntoNotes development set. CONFIDENT is the COMBINED model in which singleton is predicted if $\Pr < 0.2$ and coreferent if $\Pr > 0.8$.

model is trained on real usage data, it is not surprising that it reflects these pragmatic factors rather than just the lexical semantics (de Marneffe et al., 2012).

As noted earlier, features in Table 4 are not standardly used in coreference systems. Uryupina (2009) notes that the Karttunen features she implemented (see Section 3) do not significantly improve the performance of her discourse-new mention and non-antecedent detectors. Contrary to Uryupina, adding the features in Table 4 to a model which only incorporates the features described in Table 2 and Table 3 results in a significantly better model (likelihood ratio test, $p < 0.001$). The accuracy on the CoNLL-2012 development set also improves when adding the Karttunen features (McNemar's test, $p < 0.001$).

5.4 Results

As highlighted above, the lifespan model we built from the OntoNotes data confirms the claims by Karttunen and others concerning how semantic operators interact with specific kinds of mention. This is novel quantitative evidence for such theories. The model also successfully learns to tease singleton and coreferent mentions apart, suggesting that it has practical value for NLP applications. The first row of Table 5 summarizes the linguistic model performance on the development set of the OntoNotes data described in Section 4, giving precision, recall, and F1 measures for singleton and coreferent mentions. The accuracy of the model is 76.3%. A majority baseline, predicting all mentions as singletons, leads to an accuracy of 55.1%.

6. Extension to Bridging

The lifespan model suggests a new perspective on *bridging anaphora*, which we discussed briefly in Section 3.2 using example (14), repeated here:

(15) I looked into **the room**. *The ceiling* was very high.

The anchor phrase *the room* is superficially singleton in this discourse, but its intuitive lifespan is longer: it makes salient a discourse referent for the ceiling of *the room*, which *the ceiling* in the second sentence then refers to. The bridging relationship keeps the room alive as a discourse referent, extending its lifespan, though not in a way that can be read directly off of the text. Together with the basic tenets of the lifespan model, these observations suggest a testable hypothesis about bridging: even when bridging anchors are superficially singleton (henceforth, *singleton anchors*),[1] our lifespan model should tend to classify them as coreferent, since the model is not designed to detect later mentions per se, but rather to capture more abstract information about the roles that entities play in discourse.

OntoNotes does not contain annotations for bridging anaphora, so evaluating this hypothesis is not straightforward. However, Hou, Markert, and Strube (2013) annotated 50 of the WSJ texts in OntoNotes for bridging information, yielding annotations for 663 bridging anchors. Of these, 145 are singleton anchors in the sense that we identify them (Section 4) and thus can be used to assess our model's ability to detect the abstract sense in which bridging anchors are long-lived.

Ideally, we would simply run our trained lifespan model on these examples. This proves ineffective, though, because (outside of Hou et al.'s data) the OntoNotes annotations treat singleton anchors as singleton, meaning that our trained lifespan model is optimized on data that obscure the distinction of interest. Nonetheless, we expect the feature representations that form the backbone of the lifespan model to be able to distinguish true singletons from singleton anchors if given the right kind of training data. The small number of relevant bridging annotations poses some obstacles to pursuing this idea, but we sought to navigate around them as follows: using the annotated corpus of Hou et al., we extract all 145 of the singleton anchors and then sample an additional 145 true singletons from those documents (from a total of 5,804 such cases). This yields a data set that we can be confident makes the relevant distinction. We then randomly divide this data set into 80% training data and 20% testing data, and conduct a standard classifier evaluation. We use a logistic regression classifier, employing recursive feature elimination with cross-validation (Guyon, Weston, & Barnhill, 2002), as implemented by Pedregosa et al. (2011), to try to find a compact model that is effective for the small data set. The model used an ℓ_2 regularizer with a penalty of 0.5, though ℓ_1 regularization and changes to the penalty delivered essentially the same results, both with and without the recursive feature elimination step.

Because these train and test sets are small, performance varies greatly depending on the nature of the true singleton sample, so we repeat this experiment 1,000 times and average the results. With this procedure, our lifespan feature representations achieve a mean F1 of 65% (standard error 0.002; mean precision 62%, mean recall 0.69%), indicating our lifespan-based features are sensitive to the distinction between singleton anchors and true singletons. This finding further bolsters the design of the lifespan feature representations and also shows that "lifespan" is deeper and more abstract than merely counting referents. Given the right kind of annotations, we believe our model could be extended to provide an even fuller treatment of bridging, which is governed partly by its own mix of linguistic and contextual factors (Hawkins, 1978; Prince, 1981b; Schwarz, 2009).

7. Predicting Lifespans with Surface Features

Durrett and Klein (2013) and Hall et al. (2014) showed that, on the tasks of coreference resolution and parsing, a large quantity of surface-level information can not only implicitly model some linguistic features, but also capture other patterns in the data that are not easily identified manually. Given the large amount of annotated data available in the OntoNotes corpus, we might expect a sufficient amount of surface-level data to capture some of the linguistic insights hand-engineered in

1. Some bridging anchors also have literal coreferent mentions, as in *I looked into the room. It was empty, and the ceiling was very high.*, where *the room* is coreferent with *it* in addition to providing discourse support for *the ceiling*. We set aside such cases in our bridging experiments.

the lifespan model defined above. We therefore tested how a model using POS tags and n-grams fares on the lifespan task.

We used the following features in this surface model: the lemmas of all the words in the mention, the POS tags of all the words in the mention, the POS tag of the head of the mention, and the lemma and POS tags of the two words preceding and following the mention (with dummy BEGIN and END words to mark the beginning and end of sentences). As suggested by Durrett and Klein (2013), such features might capture information encoded in the NER tag, number, person, and sentence position.

The surface model's performance is reported in the second row of Table 5. For all models in Table 5, the ℓ_2 regularization penalty was chosen via five-fold cross-validation on the training data. For the linguistic model, using the tuned ℓ_2 regularization penalty rather than the default one makes almost no difference, but it substantially improves performance for the models with more features. We additionally experimented with different algorithms for feature selection, but found that the results were invariably best, for all our models, when we retained their full sets of features. The last row of the table gives the performance of a model in which we combine both the linguistic and the surface features to evaluate whether the surface features alone cover all the information captured by the linguistic features, or whether the linguistic features have additional predictive value.

The surface model performs better than the linguistic-only model, especially for the coreferent category. However, the small number of linguistically-motivated features yields results in the same range as those obtained with the large number of features in the surface model, which might be of importance for tasks where only a small amount of annotated data is available, such as in the bridging experiment in Section 6. (The obvious trade-off here is that the surface features are easier to specify and implement.) As shown in the COMBINED row of Table 5, combined with the surface feature set, the linguistically-motivated features give a statistically significant boost in performance. This suggests that the surface features miss certain long-distance interactions between discourse anaphora and semantic operators — interactions that the linguistic features explicitly encode.

Our best model for predicting lifespan is the combined one. Instead of using the standard 0.5 threshold as decision boundary, we can also make use of the full distribution returned by the logistic regression model and rely only on confident decisions. The resulting CONFIDENT model is a COMBINED one that predicts singleton if $\Pr < 0.2$ and coreferent if $\Pr > 0.8$. The threshold values reported here are the best trade-off we found between a precision score close to 0.90 without losing too much in recall. As expected, by using such a highly confident model, we increase precision, though at a cost to recall. Which kind of model is preferred will depend on the application; as noted by Ng (2004) and Uryupina (2009), when incorporating the lifespan model in downstream NLP applications, we often want highly accurate predictions, which favors a model like CONFIDENT.

8. Application to Coreference Resolution

To further assess the value of the lifespan model for NLP applications, we now incorporate the best feature combination into two state-of-the art coreference resolution systems: the Stanford system (Lee et al., 2011) and the Berkeley system (Durrett & Klein, 2013). In both cases, the original model serves as our baseline, and we focus on the extent to which the lifespan model contributes to improvements to that baseline. This allows us to quantify the power and effectiveness of the lifespan model in two very different systems — a rule-based one (Stanford) and a learning-based one (Berkeley).

8.1 Evaluation Measures

To evaluate the incorporation of the lifespan model into the coreference systems, we use the English development and test sets from the CoNLL-2011 and CoNLL-2012 Shared Tasks. Although the CoNLL shared tasks evaluated systems on only multi-mention (i.e., non-singleton) entities, we can still expect the lifespan model to help: by stopping singletons from being linked to multi-mention entities, we expect to see an increase in precision. Our evaluation uses the measures given by the CoNLL scorer:

- **MUC** (Vilain, Burger, Aberdeen, Connolly, & Hirschman, 1995): Link-based metric that measures how many links the gold and system partitions have in common.
- **B^3** (Bagga & Baldwin, 1998): Mention-based metric that measures the proportion of mention overlap between gold and predicted entities.
- **CEAF-ϕ_3** (Luo, 2005): Mention-based metric that, unlike B^3, enforces a one-to-one alignment between gold and predicted entities.
- **CEAF-ϕ_4** (Luo, 2005): The entity-based version of the above metric.
- **CoNLL** (Denis & Baldridge, 2009; Pradhan, Ramshaw, Marcus, Palmer, Weischedel, & Xue, 2011): Average of MUC, B^3 and CEAF-ϕ_4.
- **BLANC** (Recasens & Hovy, 2011): Link-based metric that takes the mean of coreference and non-coreference links, thereby rewarding (but not over-rewarding) singletons.

We use the new CoNLL coreference scorer (Pradhan, Luo, Recasens, Hovy, Ng, & Strube, 2014, version 8.0), which fixes a bug in previous versions concerning the way gold and predicted mentions are aligned when evaluating on automatically predicted mentions. The new scorer does not modify either the gold or system output, but implements the measures as originally proposed, and extends BLANC to successfully handle predicted mentions, following Luo, Pradhan, Recasens, and Hovy (2014).

8.2 Incorporating the Lifespan Model into the Stanford Coreference System

The Stanford system was the highest-scoring system in the CoNLL-2011 Shared Task (Pradhan et al., 2011), and was also part of the highest-scoring system (Fernandes, dos Santos, & Milidiú, 2012) in the CoNLL-2012 Shared Task (Pradhan et al., 2012). It is a rule-based system that includes a total of ten rules (or "sieves") for entity coreference, such as exact string match and pronominal resolution. The sieves are applied from highest to lowest precision, each rule adding coreference links. In each coreference resolution sieve, the document's mentions are traversed left to right. To prune the search space, if a mention has already been linked to another one by a previous sieve, only the mention that is first in textual order is considered by the subsequent sieves. Furthermore, mentions that are headed by an indefinite pronoun (e.g., *some, other*) or start with an indefinite determiner (*a, an*) are discarded if there is no antecedent that has the exact same string. Each mention is compared to the previous mentions in the text until a coreferent antecedent is found (according to the current sieve) or the beginning of the text is reached. Candidates are sorted using a left-to-right breadth-first traversal of the parse tree, which favors subjects and syntactic salience in general.

The lifespan model can improve coreference resolution in two different ways: (i) mentions classified as singletons should not be considered as either antecedents or coreferent, and (ii) mentions classified as coreferent should be linked with other mention(s). By successfully predicting singletons (i), we can enhance the system's precision; by successfully predicting coreferent mentions (ii), we can improve the system's recall. Here we focus on (i) and use the lifespan model for detecting singletons. This decision is motivated by two factors. First, given the large number of singletons (Figure 1), we are more likely to see a gain in performance from discarding singletons. Second, the multi-sieve nature of the Stanford coreference system does not make it straightforward to decide which antecedent a mention should be linked to even if we know that it is coreferent.

To integrate the singleton model into the Stanford coreference system, we depart from previous work by not letting a sieve consider whether a pair of mentions is coreferent if both mentions are classified as singletons by our CONFIDENT model and the mentions are not a named entity. In doing this, we discard 29% of the NPs under consideration. Experiments on the development set yielded higher performance when not taking into account named entities. Performance was higher with the CONFIDENT model than with the STANDARD model.

We therefore use the lifespan model to help coreference resolution as a pre-filtering step to coreference resolution, discarding mentions tagged as singletons by the lifespan model. Previous work on incorporating a non-referentiality or discourse-new detection module as a pre-processing step for coreference resolution has shown mixed results, as we discussed in Section 3. The general arguments for pipeline vs. joint approaches apply here: pipeline approaches prevent recovering from errors earlier in the pipeline, but joint approaches tend to increase model complexity and associated optimization challenges, and they do not easily allow separating different modules, which makes feature design and error analysis more difficult as well. In any case, in the context of the Stanford system's sieve-architecture, it is more natural to add the lifespan model as a pre-filtering step.

8.2.1 RESULTS

Table 6 summarizes the performance of the Stanford system on the CoNLL-2011 and CoNLL-2012 development and test sets. To evaluate the incorporation of the lifespan model in a realistic setting, we use the automatic parses, and the POS and NER tags provided in the CoNLL documents. All the scores are on automatically predicted mentions. The baseline is the Stanford coreference system, and 'w/ Lifespan' is that system extended with our lifespan model to discard singletons, as explained above. Stars indicate a statistically significant difference (Wilcoxon signed-rank test, $p < 0.05$) according to jackknifing (10 partitions of the development set or the test set, balanced over the different domains[2] of the corpus). As expected, the lifespan model significantly increases precision (up to $+4.0$ points) but decreases recall (by -0.7 points). Overall, however, the gain in precision is higher than the loss in recall, and we obtain a significant improvement of 0.4–1.5 points in the F1 score of all evaluation measures.

8.2.2 ERROR ANALYSIS

Kummerfeld and Klein (2013) provide a useful tool for automatically analyzing and categorizing errors made by coreference resolution systems. The tool identifies seven intuitive error types: span error, conflated entities (entity mentions that do not corefer are clustered together), extra entity (entities that are not in the gold data are added), extra mention (the system incorrectly introduces

2. As mentioned in Section 4, the OntoNotes corpus contains documents from seven different domains and coreference performance has been shown to vary highly depending on the domain (Pradhan et al., 2012).

Stanford	CoNLL F1	MUC R	MUC P	MUC F1	B^3 R	B^3 P	B^3 F1	CEAF-ϕ_4 R	CEAF-ϕ_4 P	CEAF-ϕ_4 F1
2011 DEV SET										
Baseline	51.49	**58.00***	55.97	56.97	**48.01***	49.81	48.89	**54.27***	44.03	48.62
w/ Lifespan	**52.23***	57.57	**57.72***	**57.65***	47.45	**51.62***	**49.45***	53.46	**46.27***	**49.60***
Discourse-new	51.52	56.30	**58.98***	**57.61***	45.51	**52.33***	48.68	48.63	**47.93***	48.28
2011 TEST SET										
Baseline	50.55	**60.09***	56.09	58.02	**47.57***	47.91	47.74	**52.28***	40.90	45.89
w/ Lifespan	**51.58***	59.75	**58.32***	**59.03***	47.06	**50.18***	**48.57***	51.42	**43.50***	**47.13***
Discourse-new	**51.26***	58.92	**59.71***	**59.31***	45.72	**51.06***	**48.25***	47.41	**45.1***	46.22
2012 DEV SET										
Baseline	55.26	**61.36***	65.26	63.25	**48.35***	57.05	52.34	**53.86***	47.01	50.20
w/ Lifespan	**55.77***	60.99	**66.70***	**63.72***	47.87	**58.57***	**52.68***	53.10	**48.91***	**50.92***
Discourse-new	53.63	60.71	63.27	61.96	47.25	54.42	50.58	49.35	**47.41***	48.36
2012 TEST SET										
Baseline	53.31	**62.05***	61.35	61.70	**48.00***	52.66	50.22	**52.29***	44.36	48.00
w/ Lifespan	**54.58***	61.31	**65.61***	**63.39***	46.91	**57.05***	**51.49***	51.03	**46.87***	**48.86***
Discourse-new	53.01	61.22	**62.73***	61.97	46.72	**53.62***	49.93	48.38	**45.92***	47.12

(a)

Stanford	CEAF-ϕ_3 R	CEAF-ϕ_3 P	CEAF-ϕ_3 F1	BLANC R	BLANC P	BLANC F1
2011 DEV SET						
Baseline	**57.11***	52.50	54.71	**45.04***	46.84	45.14
w/ Lifespan	56.55	**54.43***	**55.47***	44.37	**48.65***	**45.85***
Discourse-new	54.02	**55.67***	54.83	42.59	**49.57***	45.60
2011 TEST SET						
Baseline	**55.57***	49.56	52.39	**46.46***	47.51	46.12
w/ Lifespan	55.04	**51.80***	**53.37***	45.98	**49.53***	**47.06***
Discourse-new	53.2	**53.08***	**53.14***	44.87	**50.82***	**47.33***
2012 DEV SET						
Baseline	**56.59***	57.22	56.90	**48.78***	56.47	51.94
w/ Lifespan	56.11	**58.75***	**57.40***	48.23	**57.94***	**52.36***
Discourse-new	55.00	56.18	55.58	48.11	54.12	50.73
2012 TEST SET						
Baseline	**56.12***	53.46	54.76	**49.08***	54.48	50.88
w/ Lifespan	54.98	**56.69***	**55.82***	47.69	**59.15***	**52.28***
Discourse-new	54.43	**54.78***	54.60	47.95	**55.81***	**51.14***

(b)

Table 6: Performance of the Stanford system on the CoNLL-2011 and CoNLL-2012 development and test sets. Scores (v8.0 of the CoNLL scorer) are on automatically predicted mentions, using the CoNLL automatic annotations. Stars on the 'w/ Lifespan' and 'Discourse-new' rows indicate a significant difference from the baseline (Wilcoxon signed-rank test, $p < 0.05$).

Error	System	Gold
Conflated entities	*scientists*$_1$ *they*$_1$ *they*$_1$	*scientists*$_1$ *they*$_1$ *his family*$_2$ *they*$_2$
Extra entity	*various major Hong Kong media* *no media*	– –
Extra mention	*a book* *the book* *it*	– *the book* *it*

(a) Errors affecting precision.

Error	System	Gold
Divided entity	*scientists*$_1$ *they*$_1$ *his family*$_2$ *they*$_1$	*scientists*$_1$ *they*$_1$ *his family*$_2$ *they*$_2$
Missing entity	– –	*a network* *it*
Missing mention	*two mothers* *their* –	*two mothers* *their* *two mothers who lost very loved ones*

(b) Errors affecting recall.

Table 7: Illustration of the error types provided by Kummerfeld and Klein's (2013) system: errors made by the Stanford coreference system on the CoNLL-2012 development set.

a mention as coreferent in a cluster),[3] divided entity (an entity is split into two or more different clusters),[4] missing entity (the system fails to detect an entity), and missing mention (an entity is missing one of its mentions). Table 7 illustrates the error types we are interested in,[5] showing errors made by the Stanford system, separated into those affecting precision and those affecting recall.

We ran Kummerfeld and Klein's (2013) system on the Stanford output to quantify the improvement obtained by incorporating the lifespan model into the coreference system for the CoNLL-2012 development set. Figure 2 shows the difference in errors between the original Stanford coreference system and the system in which the lifespan model is integrated. The lifespan model generally reduces errors affecting precision, most notably by getting rid of some spurious entities ("Extra entity"). The top three errors in Table 7 — all precision-related — are fixed by integrating the lifespan model into the Stanford system. On the other hand, the bottom two errors — recall-related —

3. The distinction between the two categories *conflated entities* and *extra mention* makes sense in a corpus like OntoNotes where singletons are not annotated: the former occurs when the system clusters one or more mentions from a multi-mention entity into an incorrect entity, whereas the latter occurs when the system incorrectly clusters with others a mention that is truly part of a singleton entity (and so not annotated in the gold).

4. A *conflated-entities* error and a *divided-entity* error often co-occur.

5. The "span error" category is not relevant in the comparison here: both systems (with and without lifespan) work on the same predicted mentions.

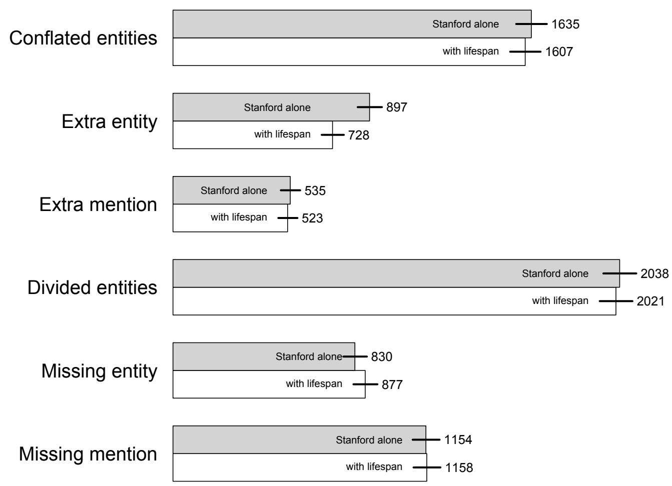

Figure 2: Number of errors for the Stanford coreference system (with and without the lifespan model) on the CoNLL-2012 development set.

are introduced by the lifespan model. However, the cumulative gain in error reduction across error categories results in a significant improvement in overall coreference performance.

8.2.3 USING THE LIFESPAN MODEL AS A DISCOURSE-NEW MENTION CLASSIFIER

As we discussed in Section 3.3, previous work (Ng & Cardie, 2002; Uryupina, 2009) reports a loss in coreference resolution performance when pre-filtering discourse-new mentions, i.e., singleton mentions as well as mentions that start a coreference chain. To mimic such pre-filtering, we incorporate the lifespan model into the Stanford system in the following way: only mentions that our model does not classify as singletons are considered by every sieve and hypothesized to corefer with some other previous mention, while discourse-new mentions are removed from consideration. When we do so, we also see a performance loss, as shown in the 'Discourse-new' rows of Table 6. There are no clear significant gains across the measures, compared to the performance of the standard Stanford system ('Baseline' rows). The improvements we do see in Table 6 result from pre-filtering pairs of mentions *both* of which our lifespan model classifies as singletons. This stricter constraint seems to balance out the loss of pre-filtering too many mentions at this early stage.

8.3 Incorporating the Lifespan Model into the Berkeley Coreference System

The Berkeley coreference system (Durrett & Klein, 2013; Durrett, Hall, & Klein, 2013) is currently the highest scoring coreference system that is publicly available. It uses a mention-synchronous framework: for each mention, the system either chooses one antecedent or decides that the mention starts a new cluster (perhaps leading to a singleton cluster). It is a log-linear model in which features are extracted over mentions to decide whether or not the mentions are anaphoric, and features are extracted over pairs of mentions to decide whether or not the pairs corefer. The baseline we compare against takes the best feature set, the 'FINAL' one, as reported by Durrett and Klein (2013), which combines a large number of lexicalized surface features as well as semantic features.

To incorporate the lifespan model into the Berkeley system, we use the probabilities of the mentions given by the lifespan model. For each pair of mentions, we add *lifespan* features by adding the lifespan probability for each mention. We also add a *singleton* feature if both mentions have a lifespan probability below 0.2, and a *coreferent* feature if both mentions have a lifespan probability above 0.8. Unlike the Stanford architecture, where exploiting the coreferent predictions is not straightforward (Section 8.2), the learning-based setup of the Berkeley system allows us to make use of the lifespan probabilities without focusing only on singleton-class prediction.

Instead of incorporating the lifespan probabilities from the lifespan model, we also tried adding to the Berkeley system all features from the lifespan model not already present in the Berkeley system (i.e., all the features in Table 3 and Table 4). However, while it did lead to significant improvements for the CoNLL 2012 development data, it did not for the CoNLL 2012 test data. Moreover, overall results were less good than when incorporating the probabilities in the manner described above.

8.3.1 RESULTS

Table 8 shows the results of the Berkeley system on the CoNLL 2011 and 2012 development and test sets. As with the Stanford system, all the scores are on automatically predicted mentions. We use the automatic POS tags, parse trees, and NER annotations provided in the CoNLL data both for training and testing. We restrict training to the training data only.[6] The baseline is the 'FINAL' Berkeley coreference system, and 'w/ Lifespan' is the same system extended with the *lifespan*, *singleton* and *coreferent* features, as explained above. Significance is computed in the same way as for the Stanford system (we created 10 partitions of the development set or the test set, balanced over the different domains of the corpus).

In the learning-based context of the Berkeley system, the lifespan model increases precision as well as recall, leading to a final improvement in the CoNLL score of 1.0 to 2.0 points. Since we use the lifespan model for predicting both singleton and coreferent mentions, we manage to improve both precision and recall. This provides additional empirical support for splitting coreference resolution into an entity-lifespan task that predicts which mentions refer to the long-lived entities in a discourse and a coreference task that focuses on establishing coreference links between these mentions.

8.3.2 ERROR ANALYSIS

Parallel to our analysis of the Stanford coreference system output, we ran Kummerfeld and Klein's (2013) system on the Berkeley output. Figure 3 shows the difference in errors between the original Berkeley coreference system ('FINAL' feature set) and that system enhanced with the lifespan model. The enhanced system commits fewer errors affecting precision (upper part of Figure 3),

6. We also tried training on the gold POS tags, parse trees, and NER annotations provided in the CoNLL data, but using the automatic annotations at test time. This does not make any difference for the original Berkeley system. When incorporating the linguistic features (either the lifespan probabilities or all features from the lifespan model not already in the Berkeley system), such a setting does lead to significant improvements over the baseline. However, improvements do not hold consistently across the development and test sets: when compared to results obtained with training on automatic annotations, training on gold improves the performance of the linguistically informed systems only for the test set.

Berkeley	CoNLL F1	MUC R	MUC P	MUC F1	B^3 R	B^3 P	B^3 F1	CEAF-ϕ_4 R	CEAF-ϕ_4 P	CEAF-ϕ_4 F1
2011 DEV SET										
Baseline	59.72	62.67	70.22	66.23	52.19	62.54	56.90	**53.77***	58.43	56.00
w/ Lifespan	**61.03***	**64.78***	**72.24***	**68.30***	**54.65***	**63.28***	**58.65***	52.89	**59.83***	56.15
2011 TEST SET										
Baseline	59.06	64.14	71.68	67.70	50.81	61.31	55.56	**51.66***	56.34	53.90
w/ Lifespan	**59.65***	**64.96***	**73.29***	**68.87***	**51.78***	**62.38***	**56.59***	49.89	**57.62***	53.48
2012 DEV SET										
Baseline	61.49	69.06	71.32	70.17	57.10	60.55	58.78	**55.20***	55.80	55.50
w/ Lifespan	**63.42***	**70.76***	**74.30***	**72.49***	**59.35***	**62.79***	**61.02***	54.74	**58.94***	**56.76***
2012 TEST SET										
Baseline	61.06	69.17	71.96	70.54	55.77	60.50	58.04	**53.82***	55.37	54.58
w/ Lifespan	**62.15***	**70.42***	**74.07***	**72.20***	**56.87***	**62.21***	**59.42***	52.64	**57.20***	54.83

(a)

Berkeley	CEAF-ϕ_3 R	CEAF-ϕ_3 P	CEAF-ϕ_3 F1	BLANC R	BLANC P	BLANC F1
2011 DEV SET						
Baseline	58.82	65.37	61.92	50.38	59.93	54.73
w/ Lifespan	**59.29***	**66.36***	**62.63***	**52.83***	**62.92***	**57.37***
2011 TEST SET						
Baseline	56.71	63.01	59.70	49.11	59.67	53.88
w/ Lifespan	56.37	**63.96***	59.93	**50.66***	**61.87***	**55.68***
2012 DEV SET						
Baseline	62.29	64.01	63.14	60.32	60.79	60.53
w/ Lifespan	62.65	**66.18***	**64.37***	**62.19***	**63.80***	**62.86***
2012 TEST SET						
Baseline	60.83	63.12	61.95	57.70	61.79	59.68
w/ Lifespan	**61.05***	**64.68***	**62.81***	**58.92***	**63.93***	**61.32***

(b)

Table 8: Performance of the Berkeley system on the CoNLL 2011 and CoNLL 2012 development and test sets. Scores (v8.0 of the CoNLL scorer) are on automatically predicted mentions, using the CoNLL automatic annotations. Stars indicate a significant difference (Wilcoxon signed-rank test, $p < 0.05$).

but not significantly for each category. However, the cumulative gains do result in a significant improvement in overall precision. Globally, the lifespan model fixes more errors than it brings in.

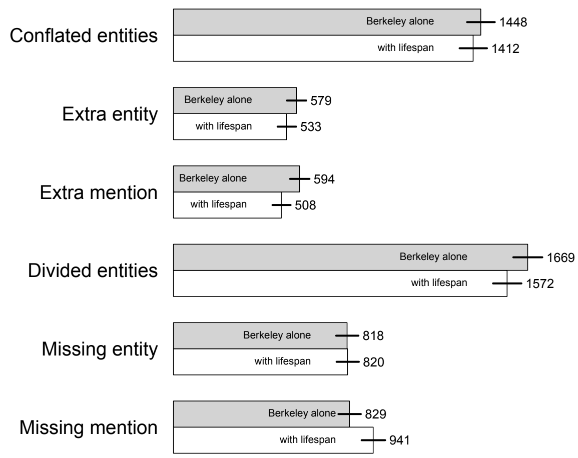

Figure 3: Number of errors for the Berkeley coreference system (with and without the lifespan model) on the CoNLL 2012 development set.

9. Conclusion

What factors determine the fate of a given discourse referent? Is it nature (its internal morphosyntax) or nurture (the broader syntactic and semantic environments of its mentions)? Our lifespan model (Section 5) suggests that nature, nurture, and their interactions are all important. The model validates existing linguistic generalizations about discourse anaphora (Section 2), and provides new insights into previous engineering efforts in a similar direction (Section 3). We also show that linguistically-motivated features bring improvement on top of surface features (Section 7), demonstrating that automatic language processing should not rely only on machine learning and big data.

The lifespan model performs well in its own right, achieving 79% accuracy in predicting whether a given mention is singleton or coreferent. This alone could have ramifications for tracking topics, identifying protagonists, and discourse coherence. In this paper, we demonstrated the benefits of the lifespan model for coreference resolution. We incorporated the lifespan model into two very different coreference resolution systems and showed that it yields improvements of practical and statistical significance in both cases (Section 8).

Stepping back, we hope to have provided a compelling illustration of how efforts in theoretical linguistics and NLP can complement each other, both for developing models and for assessing them in scientific and engineering contexts.

Acknowledgments

We thank Jefferson Barlew, Greg Durrett, Micha Elsner, Gregory Kierstead, Craige Roberts, Michael White, the Stanford NLP Group, and our anonymous reviewers for their helpful suggestions on earlier drafts of this paper. This research was supported in part by ONR grant No. N00014-10-1-0109 and ARO grant No. W911NF-07-1-0216.

References

Aissen, J. (1997). On the syntax of obviation. *Language*, *73*(4), 705–750.

Aissen, J. (2003). Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, *21*(3), 435–483.

Aloni, M. (2000). *Quantification under Conceptual Covers*. Ph.D. thesis, University of Amsterdam.

AnderBois, S., Brasoveanu, A., & Henderson, R. (2010). Crossing the appositive/at-issue meaning boundary. In Li, N., & Lutz, D. (Eds.), *Proceedings of Semantics and Linguistic Theory 20*, pp. 328–346. CLC Publications.

Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the LREC 1998 Workshop on Linguistic Coreference*, pp. 563–566.

Baker, C. L. (1970). Double negatives. *Linguistic Inquiry*, *1*(2), 169–186.

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1), 1–34.

Bean, D. L., & Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 373–380. ACL.

Beaver, D. (2004). The optimization of discourse anaphora. *Linguistics and Philosophy*, *27*(1), 3–56.

Beaver, D. I. (2007). Corpus pragmatics: Something old, something new. Paper presented at the annual meeting of the Texas Linguistic Society.

Bergsma, S., & Lin, D. (2006). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 33–40. ACL.

Bergsma, S., Lin, D., & Goebel, R. (2008). Distributional identification of non-referential pronouns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 10–18. ACL.

Bergsma, S., & Yarowsky, D. (2011). NADA: A robust system for non-referential pronoun detection. In Hendrickx, I., Lalitha Devi, S., Branco, A., & Mitkov, R. (Eds.), *Anaphora Processing and Applications*, Vol. 7099 of *Lecture Notes in Computer Science*, pp. 12–23. Springer.

Bittner, M. (2001). Surface composition as bridging. *Journal of Semantics*, *18*(2), 127–177.

Brants, T., & Franz, A. (2006). The Google Web 1T 5gram corpus version 1.1. LDC2006T13.

Byron, D. K., & Gegg-Harrison, W. (2004). Eliminating non-referring noun phrases from coreference resolution. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference*, pp. 21–26.

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, C. N. (Ed.), *Subject and Topic*, pp. 25–55. Academic Press.

Clark, H. H. (1975). Bridging. In Schank, R. C., & Nash-Webber, B. L. (Eds.), *Theoretical Issues in Natural Language Processing*, pp. 169–174. ACM.

Cresswell, M. J. (2002). Static semantics for dynamic discourse. *Linguistics and Philosophy*, *25*(5–6), 545–571.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pp. 449–454. ACL.

de Marneffe, M.-C., Manning, C. D., & Potts, C. (2012). Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, *38*(2), 301–333.

Delmonte, R., Bristot, A., Piccolino Boniforti, M. A., & Tonelli, S. (2007). Entailment and anaphora resolution in RTE3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 48–53.

Denis, P., & Baldridge, J. (2007). Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 236–243. ACL.

Denis, P., & Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, *42*, 87–96.

Durrett, G., Hall, D., & Klein, D. (2013). Decentralized entity-level modeling for coreference resolution. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 114–124. ACL.

Durrett, G., & Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1971–1982. ACL.

Elbourne, P. (2008). Demonstratives as individual concepts. *Linguistics and Philosophy*, *31*(4), 409–466.

Evans, R. (2001). Applying machine learning toward an automatic classification of “it”. *Literary and Linguistic Computing*, *16*(1), 45–57.

Fernandes, E., dos Santos, C., & Milidiú, R. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pp. 41–48. ACL.

Fodor, J. D., & Sag, I. A. (1982). Referential and quantificational indefinites. *Linguistics and Philosophy*, *5*(3), 355–398.

Fraurud, K. (1990). Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, *7*(4), 395–433.

Giampiccolo, D., Magnini, B., Dagan, I., & Dolan, B. (2007). The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9.

Groenendijk, J., & Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and Philosophy*, *14*(1), 39–100.

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

Guyon, I., Weston, J., & Barnhill, S. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422.

Hall, D., Durrett, G., & Klein, D. (2014). Less grammar, more features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 228–237. ACL.

Harris, J. A., & Potts, C. (2009). Perspective-shifting with appositives and expressives. *Linguistics and Philosophy*, *32*(6), 523–552.

Hawkins, J. A. (1978). *Definiteness and Indefiniteness*. Croom Helm.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, UMass Amherst.

Heim, I. (1992). Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, *9*(2), 183–221.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, *3*(1), 67–90.

Hou, Y., Markert, K., & Strube, M. (2013). Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 907–917. ACL.

Israel, M. (1996). Polarity sensitivity as lexical semantics. *Linguistics and Philosophy*, *19*(6), 619–666.

Israel, M. (2001). Minimizers, maximizers, and the rhetoric of scalar reasoning. *Journal of Semantics*, *18*(4), 297–331.

Israel, M. (2004). The pragmatics of polarity. In Horn, L., & Ward, G. (Eds.), *The Handbook of Pragmatics*, pp. 701–723. Blackwell.

Ji, H., & Lin, D. (2009). Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pp. 220–229.

Kamp, H. (1981). A theory of truth and discourse representation. In Groenendijk, J., Janssen, T. M. V., & Stockhof, M. (Eds.), *Formal Methods in the Study of Language*, pp. 277–322. Mathematical Centre.

Karttunen, L. (1973). Presuppositions and compound sentences. *Linguistic Inquiry*, *4*(2), 169–193.

Karttunen, L. (1976). Discourse referents. In McCawley, J. D. (Ed.), *Syntax and Semantics*, Vol. 7: Notes from the Linguistic Underground, pp. 363–385. Academic Press.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. CSLI.

Kummerfeld, J. K., & Klein, D. (2013). Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 265–277. ACL.

Ladusaw, W. A. (1996). Negation and polarity items. In Lappin, S. (Ed.), *The Handbook of Contemporary Semantic Theory*, pp. 321–341. Blackwell.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, pp. 28–34. ACL.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 25–32. ACL.

Luo, X., Pradhan, S., Recasens, M., & Hovy, E. (2014). An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 24–29. ACL.

Müller, C. (2006). Automatic detection of nonreferential *it* in spoken multi-party dialog. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pp. 49–56. ACL.

Muskens, R., van Benthem, J., & Visser (1997). Dynamics. In van Benthem, J., & ter Meulen, A. (Eds.), *Handbook of Logic and Language*, pp. 587–648. Elsevier.

Ng, V. (2004). Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 152–159. ACL.

Ng, V., & Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 1–7. ACL.

Nissim, M. (2006). Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 94–102.

Paice, C. D., & Husk, G. D. (1987). Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it". *Computer Speech & Language*, *2*(2), 109–132.

Partee, B. H. (1987). Noun phrase interpretation and type-shifting principles. In Groenendijk, J., de Jong, D., & Stokhof, M. (Eds.), *Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers*, pp. 115–143. Foris Publications.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poesio, M., Alexandrov-Kabadjov, M., Vieira, R., Goulart, R., & Uryupina, O. (2005). Does discourse-new detection help definite description resolution?. In *Proceedings of the 6th International Workshop on Computational Semantics*, pp. 236–246.

Poesio, M., Uryupina, O., Vieira, R., Alexandrov-Kabadjov, M., & Goulart, R. (2004). Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In Harabagiu, S., & Farwell, D. (Eds.), *ACL 2004: Workshop on Reference Resolution and its Applications*, pp. 47–54. ACL.

Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford University Press.

Potts, C. (2012). Conventional implicature and expressive content. In Maienborn, C., von Heusinger, K., & Portner, P. (Eds.), *Semantics: An International Handbook of Natural Language Meaning*, Vol. 3, pp. 2516–2536. Mouton de Gruyter.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., & Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 30–35. ACL. `https://github.com/conll/reference-coreference-scorers`.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pp. 1–40. ACL.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–27. ACL.

Pradhan, S. S., & Xue, N. (2009). Ontonotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pp. 11–12. ACL.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation*, pp. 2961–2968. European Language Resources Association.

Prince, E. (1981a). On the inferencing of indefinite ‘this’ NPs. In Webber, B. L., Sag, I., & Joshi, A. (Eds.), *Elements of Discourse Understanding*, pp. 231–250. Cambridge University Press.

Prince, E. F. (1981b). Toward a taxonomy of given–new information. In Cole, P. (Ed.), *Radical Pragmatics*, pp. 223–255. Academic Press.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Recasens, M., de Marneffe, M.-C., & Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 627–633. ACL.

Recasens, M., & Hovy, E. (2009). A deeper look into features for coreference resolution. In Lalitha Devi, S., Branco, A., & Mitkov, R. (Eds.), *Anaphora Processing and Applications*, Vol. 5847 of *Lecture Notes in Computer Science*, pp. 29–42. Springer.

Recasens, M., & Hovy, E. (2011). BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, *17*(4), 485–510.

Roberts, C. (1990). *Modal Subordination, Anaphora, and Distributivity*. Garland.

Roberts, C. (1996). Anaphora in intensional contexts. In Lappin, S. (Ed.), *The Handbook of Contemporary Semantic Theory*, pp. 215–246. Blackwell.

Rooryck, J. (2001). Evidentiality, Part II. *Glot International*, *5*(5), 161–168.

Saurí, R. (2008). *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

Schwarz, F. (2009). *Two Types of Definites in Natural Language*. Ph.D. thesis, UMass Amherst.

Schwarzschild, R. (2002). Singleton indefinites. *Journal of Semantics*, *19*(3), 289–314.

Simons, M. (2007). Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, *117*(6), 1034–1056.

Uryupina, O. (2003). High-precision identification of discourse new and unique noun phrases. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics Student Research Workshop*, pp. 80–86. ACL.

Uryupina, O. (2009). Detecting anaphoricity and antecedenthood for coreference resolution. *Procesamiento del lenguaje natural*, *42*, 113–120.

van Deemter, K., & Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational linguistics*, *26*(4), 629–637.

Vieira, R., & Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, *26*(4), 539–593.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*, pp. 45–52. Morgan Kaufman.

Walker, M. A., Joshi, A. K., & Prince, E. F. (Eds.). (1997). *Centering in Discourse*. Oxford University Press.

Wang, L., McCready, E., & Asher, N. (2006). Information dependency in quantificational subordination. In von Heusinger, K., & Turner, K. (Eds.), *Where Semantics Meets Pragmatics*, pp. 267–304. Elsevier.

Ward, G., & Birner, B. (2004). Information structure and non-canonical syntax. In Horn, L. R., & Ward, G. (Eds.), *The Handbook of Pragmatics*, pp. 153–174. Blackwell Publishing Ltd.