Getting to "Hearer-old": Charting Referring Expressions Across Time

Ieva Staliūnaitė[‡] **Hannah Rohde**[‡‡] **Bonnie Webber**[†] **Annie Louis**[†,¶]
[‡]Utrecht Institute of Linguistics, Utrecht University, Netherlands
[‡‡]Linguistics and English Language, University of Edinburgh, UK
[†]School of Informatics, University of Edinburgh, UK
[¶]The Alan Turing Institute, London, UK
i.r.staliunaite@students.uu.nl
{hannah.rohde, bonnie.webber, annie.louis}@ed.ac.uk

Abstract

When a reader is first introduced to an entity, its referring expression must describe the entity. For entities that are widely known, a single word or phrase often suffices. This paper presents the first study of how expressions that refer to the same entity develop over time. We track thousands of person and organization entities over 20 years of *New York Times* (NYT). As entities move from *hearer-new* (first introduction to the NYT audience) to *hearer-old* (common knowledge) status, we show empirically that the referring expressions along this trajectory depend on the type of the entity, and exhibit linguistic properties related to becoming common knowledge (e.g., shorter length, less use of appositives, more definiteness). These properties can also be used to build a model to predict how long it will take for an entity to reach hearer-old status. Our results reach 10-30% absolute improvement over a majority-class baseline.

1 Introduction

While today the company Google is so well known that its name can even be used as a verb, in 2002, it was referred to in *The New York Times*[1] as "Google, the company behind the popular Web search engine". The appositive told the readers what the company does, whereas now such elaboration is needed rarely, if at all. This paper presents a first computational study that relates the form of an entity's referring expressions (RE) in articles written at different times to the entity's changing information status.[2]

Previous work has focused on predicting how REs for an entity vary *within a single text*. This type of information status can improve coreference resolution (Recasens et al., 2013) and help generate references in automatic summaries (Nenkova and McKeown, 2003). But there has been little exploration of the change in REs to an entity *over time* and *across articles*, as the entity is accepted into common knowledge. The current work is driven by linguistic interest in characterizing REs over time. In addition, knowing the current acceptance of an entity can help in generating time-appropriate expressions. From a social science perspective, there is also great interest in capturing the birth, acceptance into common parlance, but also possible death, and subsequent reintroductions of entities.

In this paper, we disambiguate and track thousands of person (PER) and organization (ORG) entities in the *New York Times Annotated Corpus* (Sandhaus, 2008) across 20 years of news. We extract and analyze hundreds of thousands of REs for these entities, and provide the first empirical evidence that the expressions used to refer to an entity grow shorter over time, and that properties such as definiteness increase. The properties of RE form are also not uniform over entities, and we identify systematic differences between PER and ORG.

We also present a model that predicts the future information status of an entity. The model takes the REs in a small snapshot (a month's span) from anywhere in an entity's timeline, and predicts how long it will take the entity to reach hearer-old (common knowledge) status. Features related to mention frequency, and content, syntax, and topic of the REs are highly predictive, giving accuracies in the 60-80% range. We also make our corpus available for future work on REs over time.

2 Background

The choice of an RE depends on the availability and novelty of an entity (Prince, 1992). A mention may be first or later *within a text* (discourse-new

[1] article 1386221 from Sandhaus (2008)

[2] Corpus available at http://groups.inf.ed.ac.uk/cup/ref/

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4350–4359
Brussels, Belgium, October 31 - November 4, 2018.

and discourse-old respectively), or either newly introduce the entity to an audience (hearer-new) or be part of common knowledge (hearer-old). According to Prince (1992), hearer-old entities are more often mentioned with definite expressions, since the hearer can pick out the unique referent based on background knowledge. Corpus studies (Nenkova and McKeown, 2003; Yoshida, 2011) have corroborated similar trends that subsequent mentions to established entities *within a discourse* tend to be reduced or definite noun phrases.

In computational work, many studies (Nissim, 2006; Rahman and Ng, 2011; Markert et al., 2012) classify entities in a text as discourse-new/old. Here, hearer-old entities are a separate *mediated* class (not introduced in the text but which readers infer based on common knowledge), and predicted using features such as definiteness and the entity name itself. In the context of text summarization, the system of Siddharthan et al. (2011) classifies entities in source documents as hearer-old or not, based on frequency, syntax, and coreference (within the documents). The predicted status is then used in a rule-based algorithm to generate references in summaries of the source. Earlier work (Radev, 1998) modeled the choice of the best expression (from a lexicon) to fit the specific semantic context during text generation. Supplementing these efforts, we model the *progression* of entities' status from hearer-new to hearer-old as it changes across, rather than within, documents.

In the social sciences, Graus et al. (2017) analyzed distributions of entity mentions, and intervals between mentions, to identify patterns of entities becoming common knowledge, but without looking at the content of the REs. The coinage and subsequent acceptance/extinction of lexical innovations is another domain that models expressions over time, often by mapping properties of the speakers who use them within a community (see Tredici and Fernandez, 2018 and work reviewed therein). Our focus here is on REs specifically.

In what follows, we explain our RE extraction (Section 3) and linguistic features (Section 4). Analysis of the REs and the model for information status prediction are in Sections 5 and 6.

3 Extracting REs over time

We use the *New York Times Annotated Corpus* (NYTAC) (Sandhaus, 2008), containing the 1.8M articles published in the *New York Times* over the period 1987–2007 (20 years).

Given the complexity in identifying potentially interesting entities and disambiguating references to them over time, we limited our scope to person (PER) and organization (ORG) entities, which we could disambiguate with high accuracy. We also set aside years 2003–2007 for future validation, and used 1990–2002 for all our training, validation, and testing. Across these 13 years[3], we collected 52,338 unique entities (74% PER and 26% ORG) that were mentioned 284,064 times (65% PER mentions and 35% ORG).

3.1 RE span detection

We used the Stanford CoreNLP toolkit (Manning et al., 2014) (version 2018-01-31) to obtain constituency parse trees, and perform coreference resolution on the articles.[4]

We then extract all noun phrases (NP), including NPs that are nested in other NPs. For each NP, we identify if it is a *proper name* RE by matching the structure of its children to a set of six patterns (Table 1) of syntactic structures that can be used in describing an entity: *pre-modifier, relative clause, appositive, participle clause, adjective/adverb clause and prepositional phrase*.

Pattern 1 (pre-modifier) is used to identify NPs which are headed (rightmost leaf, except for possessives) by a proper noun (NNP). If an NP matches pattern 1, we have found a proper noun RE. Otherwise, if the NP matches one of the other patterns, we recursively match the patterns for the head NP phrase (bolded in Table 1) until one of them matches pattern 1. For *appositives*, there is no consensus about which NP should be the head of the phrase, as the entity name can be syntactically realized as the first or the second NP. Here we process both NPs. If there is no match for pattern 1 in this process, we discard the RE because the entity name is not a proper noun phrase. If we find multiple overlapping RE spans headed by the same noun, we only keep the largest NP.

After identifying a proper noun NP, the sequence of NNP children in the component NP span that matched pattern 1 is treated as the *base expression* (the entity name). The rest of the RE is the *descriptor* (description of the entity). For example, the phrase 'the pilot, First Lieut. Kelly Flinn' matches the *appositive* pattern.

[3] Our corpus contains trajectories for 20 years.

[4] The shift-reduce parser (Zhu et al., 2013) has an F1 score of 90% on the Wall Street Journal corpus (results from 2014).

Type	Pattern for the children of an NP	Example NP	Base expression
1. Pre-modifier	DT? (JJ|JJR|JJS|VBG|CD |QP|NP|NN|NNS|NNP| NNPS|PRP|,|CC|HYPH| SYM)* POS? *NNP*+ POS?	The tedious, complicated ABC	ABC
2. Relative clause	**NP** ,? SBAR ,?	The International Business Machines Corporation, which is the second-biggest advertiser on the Internet	International Business Machines
3. Appositive	**NP** (,|:)? **NP** (,|:)?	International Business Machines Corporation, the worlds largest computer company	International Business Machines Corporation
4. Participle clause	**NP** ,? VP ,?	International Business Machines Corporation, based in Armonk, N.Y.	International Business Machines Corporation
5. Adjective or adverb clause	**NP** ,? (ADJP|ADVP) ,?	Western Resources Inc., worth $1.7 billion	Western Resources Inc.
6. Prepositional phrase	**NP** ,? PP ,?	The National Basketball Association in New York	National Basketball Association

Table 1: Regular expressions (regex) for finding RE spans within an NP. The regex use Penn Treebank (Marcus et al., 1993) tags.

The full phrase is the RE, the string of NNPs 'First Lieut. Kelly Flinn' that comprise the embedded NP is the base expression, and the remaining NP 'the pilot' is the descriptor.

Certain adjustments had to be made. For example, NPs of the form 'NNP of NNP' (e.g. "University of Virginia") are treated as base expressions. In addition, some connectives and symbols were included in the base to accommodate names such as "Food and Drug Administration". These adjustments lead to their own errors. For instance, for "Dan Zegart of Titusville", our exception rule will mark the full expression as the base, while we would prefer just the name. As expected, our RE spans are also subject to parsing errors.

But we found that the REs mined are largely robust. We manually annotated 100 randomly selected REs for correctness of the full RE span and base expression. 92 were correctly identified. Most errors involve prepositional phrase (PP) attachment, so PP modifiers are ignored in further analysis (except PPs within base expressions).

3.2 Finding cross-document entity chains

We next identify ORG and PER entities, and link the REs for the same entities over the entire span of the NYTAC. While cross-document coreference is usually a hard problem, we took advantage of NYTAC metadata to identify and link mentions to the same ORG and PER entity with high precision.

STEP 1: Our analysis focuses on salient and repeatedly appearing entities in the news. Therefore using position in the lead paragraph as a proxy for salience, we only include entities mentioned at least once in the first three sentences of each article. We also filter out entities that appear in fewer than two documents in our corpus.

Then within an article, we extract coreference chains to identify unique entities and their REs. For each entity, we then identify a *single RE* in the article which is indicative of its hearer-old/new status. One would expect the first mention to be performing this task as later mentions are discourse-old, which itself affects the form of the REs. But we found that entities are not necessarily introduced in the first mention, hence for each entity we take the *longest descriptor* among the first three mentions in an article.

STEP 2: Next we link mentions across articles. Each NYTAC article has metadata tags which *inter alia* name *salient* people and organizations appearing in the article. These tags uniquely identify an entity every time it appears in the corpus.

We match these tags to the article REs. The tags contain normalized entity names which may not match the article's REs (from STEP 1) exactly. So we perform matches at the level of coreference chains. A chain is matched to a tagged name if the words from the base expressions in that chain overlap highly with the tag. In the case of people, the last name of the tagged person had to match the last proper noun in one of the base expressions. For ORG, at most one word in the base expression could be missing in the tag, unless an acronym of the tagged name was used as the base expression. We manually annotated the correctness of

tag matches for a random sample of 25 each of PER and ORG entities. 88% of PER and 96% of ORG matches were accurate indicating high precision. All the PER match mistakes involved two people with the same last name mentioned in same article, e.g. Bill Clinton and Hillary Clinton.[5] As an estimate of recall, we find that 61.1% of all PER metadata tags, and 59.5% of ORG tags were matched to an entity, which is reasonable. Note also that we only match when an entity is mentioned within the first three sentences of an article.

Once tags are matched to article REs, the linking of REs across the corpus is also complete since the tags are already linked. The extracted REs for each entity are ordered by the publication date of the article, creating a chain of time-ordered REs for that entity across the entire corpus.

For an example of the result of our complete timeline extraction method, consider the entity Boris Yeltsin. Mr. Yeltsin became leader of the Russian parliament in May 1990, and subsequently the President in June 1991. Both events are within the timeframe of our corpus. Below we list some expressions taken from different timepoints in Mr. Yeltsin's path to presidency and later.

Spring 1990	Mr. Yeltsin, the popular chairman of the Russian Parliament who has emerged as the champion of radical reform and decentralization and as the prime political rival to the Soviet President, Mikhail S. Gorbachev
Fall 1990	Boris Yeltsin, champion of the slender insurgent minority at the Communist Party congress
Spring 1991	Boris Yeltsin, the president of the Russian federated republic and chief opposition critic of Mr. Gorbachev
Fall 1991	Boris Yeltsin, the president of the Russian republic
Spring 1992	President Boris Yeltsin
Spring 1993	Mr. Yeltsin
Spring 2000	Former Russian President Boris Yeltsin

Figure 1 shows the average length of descriptors for Mr. Yeltsin: 5 or 6 words up until Spring 1991, after which the average length is 1 word or below, indicating a significant shift in information status.

3.3 Defining hearer-old status

Next we designate an entity's mentions as hearer-new, mentions which are hearer-old, and those whose information status lies in between.

[5]Coreference chains are also noisy in these cases.

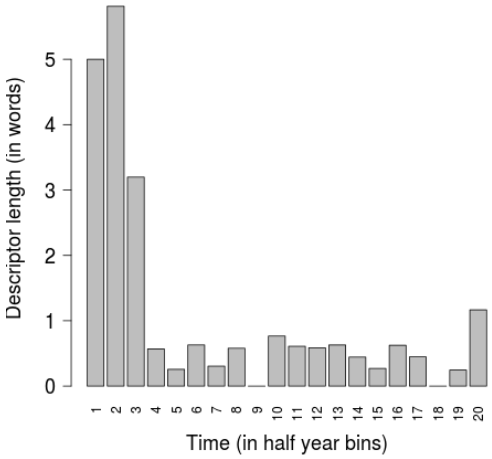


Figure 1: Descriptor lengths for Mr. Boris Yeltsin starting Spring 1990

Hearer-new mention. The NYTAC spans the years 1987–2007. To identify entities which are new, we treat the years 1987 to 1989 as a prior background corpus. When an entity is mentioned in the news from 1990 onwards but which never appeared in the prior corpus, the entity is considered as 'hearer-new' at the first mention.[6]

Mr. Yeltsin (our previous example) in fact resigned from the Politburo in 1987 before emerging again as leader of Parliament in 1990. Interestingly, our method (possibly) correctly identifies Mr. Yeltsin as hearer-new in 1990 based on comparison with the prior corpus of 1987-89.

Entity mention trajectories. From the previous section, we have the trajectories of REs for any identified hearer-new entity up to the end of our data (2002). We exclude entities that have a gap of more than 6 months between consecutive mentions, because such long gaps require the author to reintroduce an entity in case readers had forgotten it. Figure 2 shows how the length of descriptors tends to increase with greater time gaps.

Also, different entities are introduced at different times in the span from 1990 to 2002. To normalize their introduction time, we define the idea of *age* for an entity at a certain time point, so

[6]We noticed that sometimes multiple unique metadata tags exist for the same entity, mostly reflecting major changes in what the entity is known for, for example an entity before and after becoming a President. This pattern is not consistent in the editorial conventions however, so we leave improvements to cross-document coreference for future work.

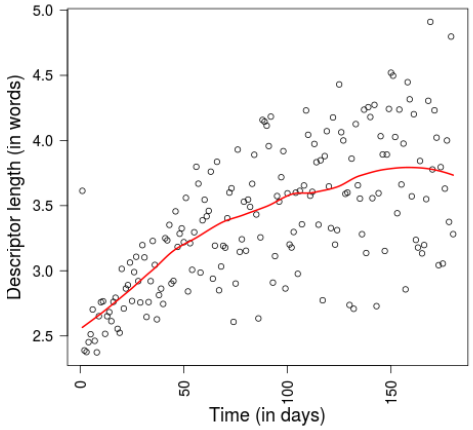


Figure 2: Mean descriptor length vs. time since previous mention

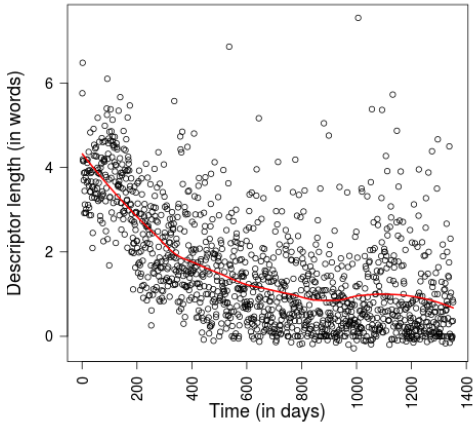


Figure 3: Mean descriptor length vs 'age' of entity

that all entities start out at time zero. For every later mention of the entity, we record the *age* as the number of days since first mention. Figure 3 shows a clear decrease in length of descriptors with increasing age of the entity.

Hearer-old mentions. We hypothesize that hearer-old entities are referred to by bare names only, and only occasionally by longer REs. Thus, we define *acceptance* of a hearer-old entity as the time after which its descriptors (the additional words besides the base expression) do not exceed a length of n words on average. Also rather than define this time point based on a single mention (which would be fragile), we bin the age values into month spans. The information status of an entity will undergo little change within a month.

First, we calculate the average length of all descriptors (entire timeline) for a given entity. If it is less than n words, then we designate the entity as hearer-old within the first month. Otherwise, we find the first month in the timeline after which the average length of descriptors of all remaining mentions is below n, and use it as time of acceptance. If the entity never reaches the threshold, we conclude that it is not accepted in the time span of the corpus. We tested values for n of 0.5 words, 1 word, and 2 words, and chose the 1-word threshold which performed best (in our linear model and classification experiments which follow).

Table 2 shows the distribution of the acceptance age for entities in our corpus. While PERs are accepted faster, more of the PER entities are also never accepted within the corpus timespan.

Below we list some examples, and qualitative observations of PER entities from each bin. The frequency of mentions is indicated within [].

1 month bin involves prominent public figures:

Bill Clinton [6005]; Michael Jordan [8]; Michael Bloomberg [546]; Nelson Mandela [3]

within 1 year are authors, journalists, local politicians (senators, mayors):

Judith Kaye (a centrist on the New York State Court of Appeals) [6]; (Senator-elect) Hillary Rodham Clinton [86]; Jurate Kazickas (a freelance writer) [2]

1-13 years involve politicians and their families, people linked to famous criminal cases:

Lisa Olson (a reporter for the Boston Herald) [14]; Laura Bush (the wife of president-elect George W. Bush) [47]; Boris Yeltsin (President of the powerful Russian Republic, whom Mr. Gorbachev portrayed as a destructive opportunist) [1229]; Joseph Gambino (a convicted heroin trafficker) [9]; Abner Louima (the Haitian immigrant who the prosecutors say was tortured by New York City police officers in a Brooklyn station house) [276]

longer than timespan involves those mentioned along a longer period but not famous (lawyers, economists, doctors), or people mentioned a lot within a short period (wedding announcements, serial killers, accident victims) and people reintroduced from the past:

Mildred Natwick (a versatile actress who created an engaging gallery of eccentric, whimsical and spunky characters in plays, films and television for more than 60 years) [2], Irene Neal (a 53 year old sculptor and painter) [2]

4 Features to characterize REs

We group the mentions of an entity within each month, and treat the *group* as one example. In this

Accepted after	less than 1 month	1 to 12 months	1 to 13 years	longer than timespan of corpus
People (entity level)	48.39%	6.42%	0.24%	44.95%
Organizations (entity level)	57.61%	8.66%	0.86%	32.89%
People (RE level)	46.92%	7.59%	2.47%	43.02%
Organizations (RE level)	49.71%	10.77%	6.67%	32.85%

Table 2: Distribution of acceptance ages. The 'entity level' rows record *one age* for each entity since its first mention. 'RE level' is the distribution when the remaining age is calculated from each RE mention. The timelines are truncated at year 2002. The later years are kept as a test set.

ORGANIZATION	**PERSON**
Positive indefinite article, length 0, length 0-3, length 3-10, length 10-20, average length, gap between mentions, topic:transport	**Positive** definite article, pre-modifier, age, length 0-3, length 3-10, length 10-20, average length, section:cars and lifestyle, topic:sports, topic: entertainment
Negative relative clause, definite article, named entity, topic:international relations, topic:military and politics	**Negative** relative clause, appositive, possessive, named entity, length 0, topic:transport, topic:religion, topic:awards

Table 3: Significant main effects in linear model

way, the example contains RE choices for the entity at *a certain snapshot in time*. A set of 61 features are computed for each example to characterize the REs (descriptor part only) it contains. *None* of the features involve the identity of the entity.

Descriptor contents: include the number of PER and ORG named entities if any in the descriptor text, type/token ratio, the counts of date and money, adjectives, superlative adjectives, verbs, and honorifics (based on a list with words such as Judge, President, Dr. etc). All counts are normalized by the number of REs in the example.

Syntactic form: They include average number per RE of definite articles, indefinite articles and possessive constructions, appositives, participle clauses, relative clauses, adjective or adverb clauses, and pre-modification. We also record the length of the descriptors using 5 bins (0 words, 1-3, 3-10, 10-20 and >20 words). The feature value is the proportion of REs in a bin. A binary feature also indicates whether the average length of descriptors is below 0.5 words.

Frequency of mentions: We include the number of mentions within the one month bin, and the average time gap between consecutive mentions (multiplied by the log of number of mentions to compensate for frequency). We also include the entity's current age (time since introduction).

Context: We employ the topic metadata from NYTAC to capture a notion of world context of an entity's mentions. Every article has topic tags (sports, finance, technology, politics, etc.) and also a section label (travel, economics, culture, etc.). We clustered the thousands of topic tags into 20 broad topics by using the Glove word embeddings (Pennington et al., 2014), and K-means clustering. A small set of 17 clusters were also created for the newspaper sections. The count of mentions belonging to articles in each cluster is a feature.

5 Features versus acceptance time

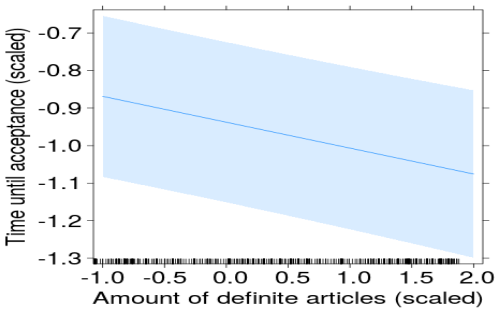
We built a linear model (LM) to test which features are significantly predictive of time to acceptance.

An example is the set of REs for an entity from a one month bin. The dependent variable is the *time left until acceptance*, i.e. the value (in months) from the current age of the entity (month were the mentions were taken from) until the acceptance age. For our corpus, the possible values are 0 to 156. Entities that do not become well known within the span of our dataset are given a label of 160 months. Entities which fall out of use (never mentioned in the corpus after a certain time point but also have not reached acceptance threshold at their found last mentions) have a label of 161.[7]

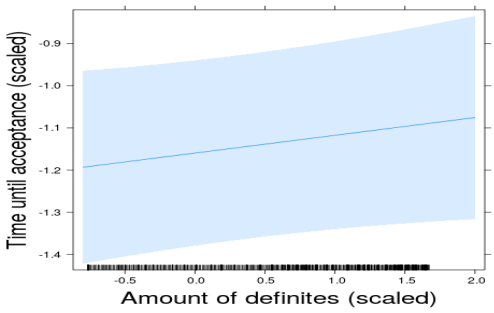
We also performed a variance analysis over the LM with ANOVA to account for possible correlation between variables. We scale the feature values into z-scores which show how many standard deviations each example is away from the mean.

We used *lm* and *Anova* (Fox and Weisberg, 2011) functions within R (R Core Team, 2013). The adjusted R^2 value is 43% for PER and 37%

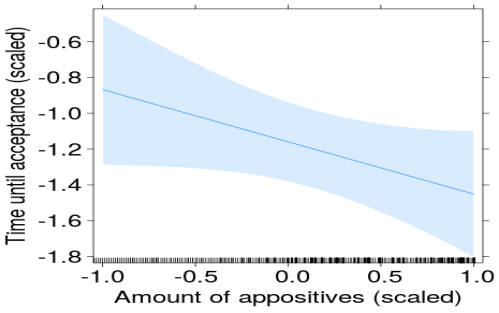
[7]One could also perform survival analysis for this problem, we have not explored it yet.



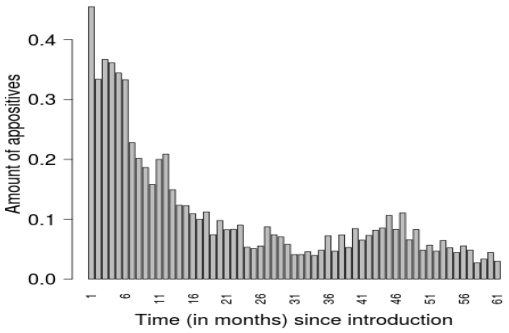
(a) Definite articles in ORG descriptors



(b) Definite articles in PER descriptors



(c) Appositives in PER descriptors



(d) Proportion of PER mentions with appositive descriptors, at different 'ages' of an entity (within the first 5 years)

Figure 4: Main effects for definite articles and appositives

for ORG entities. The significant features ($p < 0.05$) from the ANOVA are in Table 3. The features are divided into positive and negative ones, reflecting whether a higher value of the feature was predictive of a higher (positive) or lower (negative) value of the time until acceptance. More features are significant for PER (18) compared to ORG (13) expressions. Also for persons, the significant features are more varied compared to ORG, indicating that people may have more varied roles and characterizations presented in the media. Note also that we have more data for PER entities.

Below we describe some of the most interesting findings from the model.

5.1 Main effects

Definite articles are a significant feature for both PER and ORG entities. In Figures 4a and 4b, we plot the variation in *time until acceptance* versus number of definite articles, holding other variables at their means. The bands show the 0.95 level confidence interval. We used the *effects* package (Fox, 2003) in R, and the graphs show scaled values.

Assuming that the use of definite articles indicates definite expressions roughly, for ORG entities, the results confirm the hypothesis that REs become more definite the longer the entity is in use. REs such as "Ebay, an online auction site" become "Ebay, the largest of the auction sites" and "Ebay, the auction site" before settling to "Ebay".

The pattern for people is opposite, with more definite articles used with early mentions. Upon closer observation, we found that in fact, the expressions are more often definite when they are closer to acceptance, but with definiteness not expressed by the article. The definite article is excluded, and the possessive is used to produce a concise phrase e.g. "Russia's acting president" instead of "the acting president of Russia". The possessive phrase is short and also has fewer function words, which is consistent with psycholinguistic findings that people tend to reduce highly predictable phrases by dropping function words (Jaeger and Levy, 2007). Moreover, we observed that for ORG, an article may still describe the entity after acceptance e.g. "The FDA", and understandably this pattern does not exist for persons.

Appositives are not significant predictors for ORG; for PER, more appositives is associated with closeness to acceptance (Figure 4c). Yet early mentions involve more appositives than later ones (Figure 4d). One interpretation of these contrast-

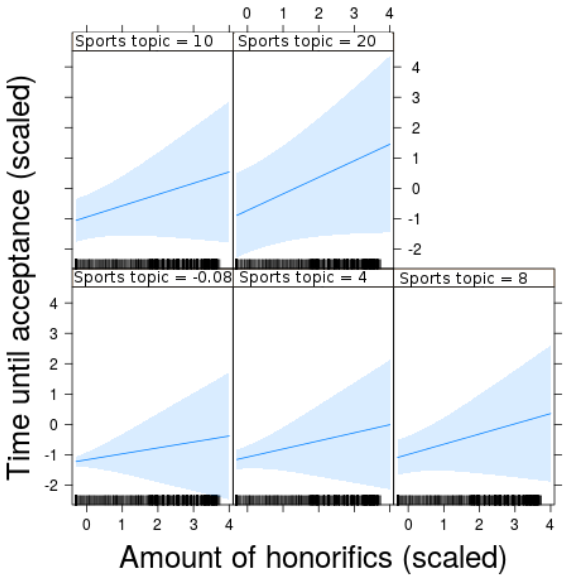


Figure 5: Interaction between the effect of sports topic and honorifics (both scaled). Sub-figures represent proportions of mentions in sports topics.

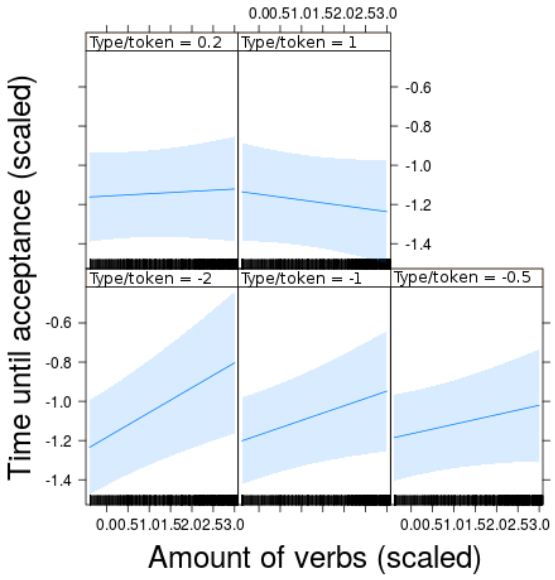


Figure 6: Interaction between the effect of type/token ratio and verb use (both scaled). Sub-figures represent values of the type/token ratio.

ing tendencies could be that there are two types of entities. The first are those important enough to be introduced with lots of appositives at the beginning of their 'lifespan'. These entities could be accepted quickly, and this behavior may manifest as having more appositives close to acceptance. The other set is that of non-salient entities which are probably not introduced with appositives anyway, and show slow acceptance. This analysis is supported by our observations that entities which are not as important as the context they appear in, such as a lesser known football player who scores a goal or an architect of a famous building, are often introduced without an appositive.

5.2 Interactions

We also found significant interactions in the model highlighting differences between various types of entities. For example, there is no main effect for honorifics but there is a significant positive effect when sports entities are involved (Figure 5). It is possible that REs such as "the number one player" indicate closeness to acceptance more than "the president of the National Hockey League", as honorifics (e.g. 'president') are not used with most known people in the sports domain.

The presence of verbs (in the descriptor) also does not have a main effect, but becomes meaningful when the type/token ratio of descriptors of a person is low (see Figure 6). We consider the type/token ratio as reflecting the different guises that a person is mentioned with, and verbs in the descriptor as associating the person with an event rather than their role in society. When the type/token ratio is low, it suggests the person is known for one or a few aspects. Here acceptance time and verbs are positively related, implying that when the few aspects are events (more verbs), the entities have a slower path to hearer-old. Whereas, if the aspects were titles such as CEO (less verbs), the entity is closer to acceptance.

6 Predicting future information-status

Given the significant correlations between our features and the time to acceptance, we now build a predictive model to identify acceptance time given a snapshot of REs for an entity. Every example contains a certain entity's mentions over a month's span, similar to the linear model. The target class indicates whether or not the entity was accepted within a period, x years from the sample time.

The classification models were built separately for PER and ORG expressions. We divide the data into 70% training, 10% validation and 20% test. For number of data points for training/validation/test, ORG has 8,768/1,252/2,506 data points; PER has 26,190/3,742/7,483. Note that examples are month-sized bins of each en-

ENTITY	Baseline (%) (majority-class)	SVM			MLP			
		C	Gamma	Accuracy (%)	Layers	Units	Alpha	Accuracy (%)
4-CLASS								
Org	35.17	10	0.01	62.69	2	400	100	68.36
People	42.01	100	0.01	70.77	3	300	100	74.18
BINARY								
Org	69.80	10	0.01	78.41	3	100	100	79.41
People	59.16	10	0.01	79.14	3	300	100	80.66

Table 4: Results of the SVM and MLP classifiers. C and Alpha are regularization parameters, Gamma parameterizes the RBF kernel. Hidden layers and hidden units in the MLP are also shown.

tity's REs, rather than unique entities. We explore both an SVM classifier with an RBF kernel, and a Multi Layer Perceptron (MLP) classifier. The MLP uses a BFGS solver with ReLU activation function. We used implementations from scikit-learn (Pedregosa et al., 2011), and the parameters of both classifiers were tuned on the development set using grid search.

We build two types of classifiers (results in Table 4). The baseline is majority-class assignment.

4-CLASS is a four-label classifier for predicting *when* an entity will become hearer-old. The four classes reflect the distribution of acceptance ages (see Table 2—RE level): already accepted (at the sample month), will be accepted within the year, between 1 and 13 years from the sample, and will not be accepted within the time frame we have.

BINARY is a classifier performing a simpler binary division of whether an entity is hearer-old or hearer-new *after 2 years* from the sampling time. About half of both entities in our data are accepted within 2 years after introduction.

The 4-CLASS MLP model is better than SVM reaching 68% accuracy for ORG and 74% for people. The improvement is over 30% absolute value, indicating the significant effect of the model and features. Still there is scope for improvement, given that the performance is less than 75%.

Both SVM and MLP perform similarly for the binary tasks. The overall binary classification accuracy is 80% for both PER and ORG entities, a 10% increase for ORG and 20% for PER.

We also performed an ablation study to identity the most useful individual classes of features. Syntax features (Section 4) had the biggest impact when removed, lowering performing by 5-8% for all models. But since both our classifiers are non-linear, they can capture useful interactions between all our feature classes.[8]

[8]The *context* class uses corpus-specific metadata, but unsupervised topic modeling could likely approximate it.

7 Conclusion

In this work, we have shown empirically that the path to hearer-old status displays detectable and interesting linguistic features, and that entities of a certain type exhibit distinctive properties. These significant differences have allowed for a first model which predicts how long it will take for an entity to be accepted as common knowledge.

There are a number of directions which we plan to explore. While we have focused on predicting whether an entity will be accepted after a certain period of time has passed, we have not modeled the RE tokens themselves, or their generation. During our analysis, it was clear that the REs of a (named) entity may also change semantically over time, reflecting current interest in it, rather than the same RE content just growing shorter over time. We have also not explored the RE timelines of common noun entities such as *organic food* or *cryptocurrency*. It is possible that they follow a different trend than named entities, and require a different set of feature indicators. We also plan to improve upon our current models and assumptions. Currently, we have ignored entities with long time gaps and requiring reintroduction to an audience. Learning to predict when an entity will go out of use, and indicators for reintroduction will add strength to our analyses. The definition of acceptance and fine-grained models for prediction will also be developed. We are also releasing a corpus of chained REs from the NYTAC (represented as byte-span sets) to enable other researchers to study these aspects of REs.

Acknowledgments

Staliūnaitė was funded under EU COST Action IS1312 (*TextLink*). Rohde was supported by a Leverhulme Trust Prize in Languages & Literatures. We also thank Jan Odijk, Mark Steedman, and our anonymous reviewers for their comments.

References

John Fox. 2003. Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15):1–27.

John Fox and Sanford Weisberg. 2011. *An R Companion to Applied Regression*, second edition. Sage, Thousand Oaks CA.

David Graus, Daan Odijk, and Maarten de Rijke. 2017. The birth of collective memories: Analyzing emerging entities in text streams. *arXiv preprint arXiv:1701.04039*.

Florian T. Jaeger and Roger P. Levy. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of Advances in Neural Information Processing Systems*, pages 849–856.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 795–804.

Ani Nenkova and Kathleen McKeown. 2003. References to named entities: A corpus study. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Short Papers - Volume 2*, pages 70–72.

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

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Ellen F Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. *In Sandra Thompson and William Mann, editors, Discourse description: diverse analyses of a fund raising text*, pages 295–325.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Dragomir R Radev. 1998. Learning correlations between linguistic indicators and semantic constraints: Reuse of context-dependent descriptions of entities. In *Proceedings of the 17th International Conference on Computational linguistics-Volume 2*, pages 1072–1078.

Altaf Rahman and Vincent Ng. 2011. Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1069–1080.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 627–633.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Marco Del Tredici and Raquel Fernandez. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Etsuko Yoshida. 2011. *Referring expressions in English and Japanese: Patterns of use in dialogue processing*, volume 208. John Benjamins Publishing.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 434–443.