A Computational Model of Human Preferences for Pronoun Resolution

Olga Seminck
LLF (CNRS)
8 Place Paul-Ricoeur, 75013 Paris
Université Paris Diderot, Paris 7
5 rue Thomas Mann, 75013 Paris
olga.seminck@cri-paris.org

Pascal Amsili
LLF (CNRS)
8 Place Paul-Ricoeur, 75013 Paris
Université Paris Diderot, Paris 7
5 rue Thomas Mann, 75013 Paris
pascal.amsili@linguist.univ-paris-diderot.fr

Abstract

We present a cognitive computational model of pronoun resolution that reproduces the human interpretation preferences of the Subject Assignment Strategy and the Parallel Function Strategy. Our model relies on a probabilistic pronoun resolution system trained on corpus data. Factors influencing pronoun resolution are represented as features weighted by their relative importance. The importance the model gives to the preferences is in line with psycholinguistic studies. We demonstrate the cognitive plausibility of the model by running it on experimental items and simulating antecedent choice and reading times of human participants. Our model can be used as a new means to study pronoun resolution, because it captures the interaction of preferences.

1 Introduction

Pronoun resolution has been studied in the frame of theories of formal grammar, corpus studies, experimental psycholinguistic studies and NLP systems.[1] But much of the findings made about the phenomenon are not shared between these disciplines. This paper takes a step towards more interdisciplinarity between the fields of NLP and psycholinguistics by building a cognitive computational model of pronoun resolution. As Keller (2010) argues convincingly, both the domains of NLP and psycholinguistics can benefit from such models. On the one hand, there is a very rich psycholinguistic literature of which researchers in the domain of NLP are often not aware. NLP techniques might improve if this literature is taken into account. On the other hand, cognitive computational models are a new means to perform psycholinguistic research: by implementing different models that represent different theories, a comparison can be made by looking at the behavior of the models on actual experimental human data.

On the topic of pronoun resolution some cognitive computational models have already been proposed. Frank et al. (2007) proposed a model that resolves ambiguous pronouns based on human interpretation biases (preferences) — such as the *first mention bias*[2]— and world knowledge. They used a so-called *micro-world*: a collection of very detailed world knowledge for a small set of events. Their model was able to simulate reading times, but it remains an open question to what extent the model can be scaled up (Frank et al., 2007).

Kehler and Rohde (2013) proposed a probabilistic model to predict human interpretation biases. Their model, based on world knowledge and information structure, predicts the probability that a given referent is mentioned next. They tested the model on human data from completion tasks[3] and showed that the model could accurately predict the human data.

Dubey et al. (2013) developed a model based on surprisal. Surprisal is a measure that is high when infrequent, or unexpected, events happen. According to Surprisal Theory (Hale, 2001), the surprisal of syntactic structures reflects their cognitive processing cost. That is to say that infrequent syntactic structures are more difficult to process for humans than frequent ones. Demberg and Keller (2008) showed that syntactic surprisal is a relevant factor to model reading times on corpus. In the model of Dubey et al. (2013) syntactic surprisal

[1] In the latter domain nowadays mostly in the form of the coreference resolution task, of which proper pronoun resolution is only a part.

[2] A character that is named first in the sentence is the preferred interpretation of ambiguous pronouns.

[3] In a completion task participants have to complete a text of which only the beginning is given.

Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 53–63, Valencia, Spain, April 3-7 2017. ©2017 Association for Computational Linguistics

is enriched by surprisal coming from coreference. Surprisal is higher when a new referent is introduced and lower when an old one is re-mentioned. Dubey et al. (2013) show that their enriched measure of surprisal is better in explaining the variance in reading times recorded on corpus than a standard measure of only syntactic surprisal.

Inspired by Dubey et al. (2013), we aim for a model of pronoun resolution that can run on natural texts and explain reading times. A second aim for our model is that it can account for human preferences discovered in the psycholinguistic literature. Based on these criteria, we build a model inspired by NLP pronoun resolution systems (Soon et al., 2001). The factors of influence on pronoun resolution are represented as weighted features. This provides a way to assess their relative importance and allows to study their interaction.

In this paper we demonstrate our model by running it on items used in psycholinguistic experiments about human preferences. We first show that the strength of human preferences corresponds to the weights our model associates to different factors influencing pronoun resolution. Second, we study how the model chooses antecedents for pronouns and see that it makes choices similar to humans. Finally, we simulate reading times by formulating a metric of processing cost based on our model.

2 Preferences Modeled in This Work

We chose to model two preferences that operate in English in this work: the Subject Assignment Strategy and the Parallel Function Strategy. We made this choice because of the feasibility of the implementation: both preferences rely only on syntactic mechanisms, so no semantic representation needed to be implemented.

The Subject Assignment Strategy states that, if a pronoun is ambiguous (*i.e.* has more than one antecedent candidate compatible in gender and number), it will be resolved to the antecedent candidate that is in the subject position (Crawley et al., 1990). So for both of the following examples the Subject Assignment Strategy predicts that the antecedent of the pronoun is *John*.

(1) a. John hit Fred and [he]$_{resolve}$ kicked Ellen.
b. John hit Fred and Ellen kicked [him]$_{resolve}$.

According to the Parallel Function Strategy, an ambiguous pronoun is resolved to the antecedent candidate that has the same syntactic function (Smyth, 1994). So according to this second strategy, in example (1-a) *he* will be resolved to *John*, whereas in (1-b) *him* will be resolved to *Fred*.

Evidence for both the Subject Assignment Strategy and the Parallel Function Strategy is not new and comes from early studies from the 1970's (Hobbs, 1976; Sheldon, 1974, among others). However, the interaction between both strategies was investigated more recently by Crawley et al. (1990). They performed two experiments with stimuli like the one in (2), where an ambiguous pronoun in the direct or indirect object position had to be resolved to either a character in the subject position (Brenda) or a character in the object position (Harriet). They chose not to study pronouns occupying the subject position, because both the Subject Assignment Strategy and the Parallel Function Strategy make the same predictions for these pronouns. Instead, they studied resolution of ambiguous pronouns in the direct and indirect object function to see the influence of both the Subject Assignment Strategy and the Parallel Function Strategy.

(2) Brenda and Harriet were starring in the local musical. Bill was in it too and none of them were very sure of their lines or the dance steps. Brenda copied Harriet and Bill watched [her]$_{resolve}$.

They found that only the Subject Assignment Strategy was used in pronoun resolution. However, different studies that followed up their paper, such as Smyth (1994) and Stevenson et al. (1995), found strong evidence for the existence of the Parallel Function Strategy alongside the Subject Assignment Strategy. They criticized the fact that many items used by Crawley et al. (1990) weren't exactly parallel: in many items none of the potential antecedents occupied exactly the same syntactic function as the pronoun. For example in item (3) there is no antecedent candidate in the direct object position (*Monica* is in an indirect object position).

(3) Cheryl and Monica were members of the local peace group. Steven had just joined and wasn't very involved yet. Cheryl spoke to Monica about the next meeting and Steven questioned [her]$_{resolve}$ about it.

With new experiments, Smyth (1994) and Stevenson et al. (1995) established the influence of the Parallel Function Strategy. They even suggested that it overrules the Subject Assignment Strategy

if it can be applied.

In our study we build a model of pronoun resolution that can account for some of the findings of Crawley et al. (1990) and of Smyth (1994). More precisely, we run our model on the items of Crawley's experiment and of Smyth's second experiment. [4]

3 Model of Pronoun Resolution

We used a classifier that proceeds according to a probabilistic version of the pair-wise algorithm (Soon et al., 2001). We only account for third person singular personal pronoun resolution in order to approach the psycholinguistic domain where pronoun resolution is most often restricted to these type of pronouns. The third person pronouns can be viewed as different from the first and the second as the latter are deictic rather than anaphorical.

3.1 Resolver

The pairwise resolver is a logistic regression classifier that gives the probability that a pair of a pronoun and an antecedent candidate are coreferent. We chose it for its straightforward interpretation of feature weights, indicating the influence of factors in pronoun resolution. We trained it on examples of pairs of coreferent and non-coreferent mentions. We used the method of Soon et al. (2001) to sample training examples: to get positive training examples (coreferent pairs), each pronoun is coupled to its closest antecedent. To get negative training examples, the pronoun forms a pair with every mention occurring between its closest antecedent and itself.

3.2 Corpus

We trained the resolver on the English newswire part of the Ontonotes 5.0 corpus (Pradhan et al., 2011). This genre approximated the psycholinguistic items the best among the available genres in Ontonotes. A particularity of the corpus is that singleton mentions (referential expressions that are only mentioned once) are not annotated. We resolved this problem by simply considering as a singleton mention every maximal noun phrase that did not overlap with an annotated mention and that was not a pronoun. Moreover, since Ontonotes is not annotated for number nor gender, we had to add (automatically) an annotation for number and gender to the mentions in the corpus. [5]

3.3 Features

The aim of our model is to have interpretable features and not to have the best score on a pronoun resolution task. We proceeded in three steps to establish the features of our classifier. First, we defined a list of standard features for pronoun resolution — inspired by coreference resolution literature (Denis and Baldridge, 2007; Recasens and Hovy, 2009; Soon et al., 2001; Yang et al., 2004) — that we could retrieve in our corpus. [6] It is important to point out that, although we made up our feature list by looking at literature from Natural Language Processing, the features in the list are also discussed in psycholinguistic literature. For example, distance features and part of speech features are discussed in literature about antecedent *saliency* (Ariel, 1991).

Among all the features, we made sure we included the features necessary to test the two preferences investigated in this paper. For the Subject Assignment Strategy, we used a feature that checks whether the antecedent candidate is in the subject position. We implemented the Parallel Function Strategy by a boolean feature of *syntactic path match* that states whether the antecedent candidate and the pronoun have the same path in the syntactic parse tree from the node where the mention is attached to the root of the tree. A simple illustration of this is given in Figure 1 where the syntactic paths of two mentions are given.

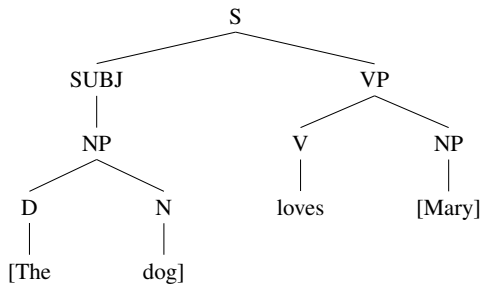


Figure 1: A syntactic tree with two mentions: *the dog* and *Mary*. Syntactic path for *the dog*: [SUBJ, S]. Syntactic path for *Mary*: [VP, S].

[4] We chose these experiments because in the remaining experiments of Smyth (1994), and also in the experiments of Stevenson et al. (1995), a different definition of the Parallel Function Strategy has been used.

[5] The procedure of gender/number annotation we chose is explained in section B of the supplementary materials.

[6] A list of these features can be found in section A of the supplementary materials.

The second step of defining our features consisted in eliminating features too sparsely represented in our training corpus to be adequately learned. As a rule of the thumb we decided to exclude features with a frequency smaller than 0.5%, meaning that every feature should be attested at least 36 times in the training data.

As a last step we checked the significance of our features and removed features that were not significant, because their interpretation is difficult. The model with the features we selected can be found in Table 1.

	Estimate	Signif.
(Intercept)	-2.3533	***
match in gender	2.4206	***
match in number	0.2430	*
m_1 is a subject	1.5142	***
match in syntactic path	1.7318	***
m_1 is a proper noun	0.5007	***
m_1 is a possessive pronoun	1.9037	***
m_1 is a personal pronoun	0.7647	***
words between m_1 and m_2	-0.0114	***
m_1 & m_2 in the same sentence	0.3587	***
length of syntactic path m_1	-0.1361	***
m_1 is determined	-0.2825	*
m_1 is undetermined	-0.4422	**
m_1 has a demonstrative determiner	0.6045	*
m_1 is a common noun	-0.8967	***
m_1 spans m_2	-3.4372	***
length in words of m_1	-0.0201	*
m_1 is a geopolitical entity	-1.2885	***
m_1 is a date	-1.9416	***

Table 1: The selected model of the pronoun resolver. Each factor influencing pronoun resolution has an estimated weight associated that indicates its influence. m_1 refers to the antecedent candidate, m_2 to the pronoun. Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.

3.4 Evaluation

We divided the corpus into a training set, a development set and a test set. We tested the model's performance on all three of the sets by measuring the accuracy of the identification of antecedents of the third person singular personal pronouns in the corpus. The accuracy and size for each subcorpus can be found in Table 2.

An important question is whether these results are satisfactory. Our results are difficult to compare against state-of-the-art work in coreference resolution, because we concentrate on third person personal singular pronouns only. This means that our system does not form coreference chains and that its performance cannot be measured using standard coreference evaluation metrics, such as MUC, B3, or CEAF (Luo, 2005). A second difference with a more standard approach is that we do not have a module of mention detection. Instead, we use the gold mention annotation and the singleton mentions we extracted (see section 3.2).

Sub-Corpus	Nb. Texts	Nb. Pronouns	Accuracy
Training	476 (60%)	1756	61.15
Development	158 (20%)	558	65.41
Test	158 (20%)	617	61.26

Table 2: The accuracy of the resolver for finding the correct antecedent of the pronoun on the training, development and test set.

This said, we still want to have an indication about the performance of our classifier. The study of Yang et al. (2004) is the most comparable we found to ours, although they used a module for mention detection. Yang et al. (2004) trained different types of systems to perform third person pronoun resolution and reported accuracy, in their paper indicated by the metric of *success*. When they tested on the MUC-6 corpus this metric was between 70.0 and 74.7 for the different systems they developed. When tested on the MUC-7 corpus the metric laid between 53.8 and 62.5. We estimate that, given these numbers, the performance of our model is slightly worse, or comparable.

An error analysis we conducted indicated that most of the errors made by the resolver concerned the pronoun 'it' (about half of the errors). We observed that if we excluded 'it' from resolution the pronoun resolver's accuracy increased by ≈ 16 points. Our error analysis also indicated that a part of the errors comes from our automatic gender annotation: it seems that many coreference chains contain mentions of several genders at once. Nevertheless, we think that the performance on masculine and feminine pronouns of our system is good enough for the purpose of our experiments that include only masculine and feminine pronouns.

3.5 Interpretation of the Model

The weights of the logistic regression model in Table 1 predict the preferences the classifier will show on experimental data. Looking at the feature of syntactic path match and the feature that checks if the first mention is in the subject position, we see that both features have a positive weight; but we can also see that the first is stronger than the second, suggesting that parallel roles are of a greater

impact than the subject position of the antecedent. From this data we can hypothesize that the Subject Assignment Strategy exists alongside the Parallel Function Strategy, and that the Parallel Function Strategy, if applicable, has a stronger influence that can overrule the Subject Assignment Strategy.

4 Antecedent Choice for Pronouns

To test the cognitive plausibility of our model, we ran it on the experimental items of Crawley et al. (1990) and the items of the second experiment of Smyth (1994) and looked if it chose the same antecedents as humans did. That is to say that we compared the model's frequencies of assigning pronouns to subjects and objects with human frequencies.

4.1 Items

For each type of item we give two examples to illustrate the type of experimental items used. Before running the model, we manually annotated the items with coreference and named entity information. For the syntactic annotation we first ran the Stanford Parser (Klein and Manning, 2003) and then corrected the parses manually.

4.1.1 Crawley's Ambiguous Items

From the experiment of Crawley et al. (1990) we have 40 ambiguous items. Ambiguity is produced by gender. The pronoun that has to be resolved is presented in the last sentence in the direct or indirect object position.

1. John and Sammy were playing in the garden. One of their classmates, Evelyn, tried to join in their game. John pushed Sammy and Evelyn kicked him.

2. Mary and Julie were about to go into town when they realized the car had a puncture. Their next door neighbour, Peter, was working in the garden. Mary helped Julie change the wheel and Peter talked to her.

4.1.2 Crawley's Unambiguous Items with Subject Antecedent

The ambiguous items have unambiguous versions: there is only one possible antecedent that matches in gender. All 40 ambiguous items (see section 4.1.1) have an unambiguous version in which the antecedent of the pronoun is the subject of the sentence in which the pronoun appears. Note that the pronoun is still always in the direct or indirect object position.

1. John and Mary were playing in the garden. One of their classmates, Evelyn, tried to join in their game. John pushed Mary and Evelyn kicked him.

2. Mary and Tim were about to go into town when they realised the car had a puncture. Their next door neighbour, Peter, was working in the garden. Mary helped Tim change the wheel and Peter talked to her.

4.1.3 Crawley's Unambiguous Items with Object Antecedent

All 40 ambiguous items from section 4.1.1 also have an ambiguous version in which the pronoun's antecedent appears at the direct or indirect object position.

1. Mary and John were playing in the garden. One of their classmates, Evelyn, tried to join in their game. Mary pushed John and Evelyn kicked him.

2. Tim and Mary were about to go into town when they realised the car had a puncture. Their next door neighbour, Peter, was working in the garden. Tim helped Mary change the wheel and Peter talked to her.

4.1.4 Smyth's Ambiguous Pronouns in Subject Position

In Smyth (1994)'s second experiment, there are ten ambiguous items with a pronoun in the subject position. A full parallelism can be found between the subject of the item and the pronoun.

1. Mary helped Julie change the tire and then she helped Peter change the oil.

2. Shirley wrote to Carol about a meeting and then she wrote to Martin about a party.

4.1.5 Smyth's Ambiguous Pronouns in Object Position

Smyth (1994) also presents ten items with a pronoun in the direct or indirect object position. For all ten items a full parallelism can be found between the pronoun and a character in the direct or indirect object position.

1. John pushed Sammy and then Evelyn kicked him.

2. Sarah visited Cathy at home and then Charles phoned her at work.

4.2 Results

We can see in Table 3 that the model fits human preferences quite accurately. With the ambiguous items from Crawley et al. (1990) we observed that the Subject Assignment Strategy applies as a default strategy when the Parallel Function Strategy is not available. For the unambiguous items, Crawley et al. (1990) did not report human assignment. The model's assignment for these items was a 100% correct when the antecedent was a subject, but when it was an object or indirect object in

Experiment	Human		Model	
	% Sub.	% Obj.	% Sub.	% Obj.
Crawley, ambiguous items, pronoun in the object position (4.1.1)	60%	40%	72.5%	27.5%
Crawley, unambiguous items, antecedent in the subject position (4.1.2)	n.a.	n.a.	100%	0%
Crawley, unambiguous items, antecedent in the object position (4.1.3)	n.a.	n.a.	0%	85%
Smyth exp. 2, ambiguous items, pronoun in the subject position (4.1.4)	100%	0%	100%	0%
Smyth exp. 2, ambiguous items, pronoun in the object position (4.1.5)	12%	88%	30%	70%

Table 3: Human pronoun assignment versus the model's predictions on Crawley et al. (1990)'s items and Smyth (1994)'s items from experiment 2. For each item set examples can be found in section 4.1. For Crawley et al. (1990)'s unambiguous items, no human results were reported. Note that for the unambiguous items with pronouns in the object position, the model sometimes did not assign any antecedent to the pronoun.

15% of the cases the model could not attribute a score high enough to choose it as the antecedent and responded *None*[7]. For the items of Smyth (1994)'s experiment, we observed — just like him — that the Parallel Function Strategy is the preferred strategy.

4.3 Discussion

We have shown that our model is able to mirror quite accurately pronoun resolution preferences. As our model is trained on real corpus data, this means that such preferences are somehow statistically presented in the language. Our model is in line with the claim that the Parallel Function Strategy and the Subject Assignment Strategy exist alongside each other and that the former can overrule the latter. Our model embodies the idea Smyth (1994) has about pronoun resolution:

> "Pronoun resolution is a feature-match process whereby the best antecedent is that which shares the most features with the pronoun."

It also captures Smyth (1994)'s idea that not every feature has the same impact and that for example *gender match* is more important than parallel roles. Based on the results our model obtains on the experimental items, we conclude that the weights it learned from corpus are cognitively plausible.

5 Simulation of Reading Times

We use our model to simulate reading times recorded in pronoun resolution experiments. An important question is: how can our model account for those reading times? It is commonly assumed that reading time is determined by the difficulty of language processing: more difficulty will result in a longer reading time. Therefore, we need a measure of 'difficulty' from our model to simulate it. We call this measure a cost metric. In the following subsection we explain how our model can output a cost metric for pronoun resolution. We then compare our metric to reading times recorded in Crawley et al. (1990)'s experiment. [8]

5.1 Cost Metric for Pronoun Resolution

To formulate a cost metric, we have to determine first what would cause cost in pronoun resolution. We hypothesize that the difficulty of finding the antecedent is determined by the number of compatible candidates and their degree of compatibility. A higher number of compatible candidates and a higher degree of compatibility will create more competition and therefore more processing cost.

Our model is able to measure compatibility of antecedents by giving a probability score to the antecedent candidates. Nevertheless, these scores do not reflect directly the competition amongst the candidates, because the resolver makes no statements about the relation between the different scores. Therefore, to measure competition, we use the notion of entropy from Information Theory (Shannon and Weaver, 1949). Entropy is a property of a random variable and captures how much uncertainty plays a role in it. The formula of entropy — in which X is a random variable that can take the values of i — is:

$$H(X) = -\sum_{i \in X} p(X = i) \cdot \log_2(p(X = i)) \qquad (1)$$

[7] Among all antecedent candidates the correct antecedent got still the highest score, but it was lower than 50%, so the resolver responded that it did not find the antecedent. This behavior of the system can been seen as the result of training it on the Ontonotes corpus, where the bias towards classifying negative must be high, to prevent it from linking pronouns to wrong antecedents.

[8] Unfortunately, in Smyth (1994)'s experiment, no measure of processing cost was taken, so we could not apply our cost metric on its experimental items.

By defining our cost metric as the entropy over the probability distribution of antecedent candidates, we can capture the idea of competition. But a problem is that a probability distribution over antecedent candidates does not follow naturally from our model. Hence, we decided to form a probability distribution from the scores we have by using techniques inspired by Luo et al. (2004), who investigated how to form a probability distribution on entities (coreference chains) by using a probabilistic mention-pair classifier, similar to our resolver. To calculate the processing cost for a pronoun, we used the following steps:

We first get from our resolver the coreference scores between every preceding mention in the text and the pronoun to be resolved. We then group the preceding mentions by their coreference chain. Because our resolution system does not build coreference chains, this information is taken from the corpus annotation.[9] As in the work of Luo et al. (2004), each chain gets the score of its highest scoring mention. Then, among the antecedent candidates, we consider all the chains that obtain a score >0.5 [10] together with an 'empty' candidate (*i.e.* the pronoun has no antecedent) in the case that the pronoun is not anaphoric, but cataphoric.[11] We also followed Luo et al. (2004) in the assignment of probability to the empty candidate: it is given a probability equal to 1 minus the score of the highest scored mention. Next, to form a probability distribution over the mentions, we used the technique described in Luo et al. (2004): a probability distribution over the chains is formed by dividing the probability for each chain by the probability mass of all the chains in the distribution. Finally, the entropy is calculated on this distribution. This procedure is illustrated in Table 4.

m_i	P(m_i)		c_i	P(c_i)	P(dist)	Entropy
box	0.95	}	{*box, its*}	0.95	0.56	1.15
its	0.85	}				
cat	0.7	}	{*cat, it*}	0.7	0.41	
it	0.6	}				
Bob	0.01	}	{*Bob, he*}	0.2	–	
he	0.2	}				
$\emptyset$	0.05	}	{$\emptyset$}	0.05	0.03	

Table 4: Imagine that in a text the pronoun *it* has to be resolved and that all preceding mentions in the text are reported under m_i. First *P(m_i)* is outputted by the resolver and indicates the probability that m_i is coreferent with *it*. The empty candidate gets the score of 1 minus the highest scoring mention (hereunder: 1 - 0.95 = 0.05). Second, each mention is associated to its corefference chain c_i. Each chain gets the probability of its highest scoring mention, reported under P(c_i). Third, a probability distribution is forged from all candidates having a *P(c_i)* > 0.5 and the empty candidate. This is done by dividing the scores under *P(c_i)* by the total probability mass of the maintained candidates (hereunder: 0.95 + 0.7 + 0.05). The result is a probability distribution, reported as *P(dist)*. Entropy is calculated on this distribution.

5.2 Results

Our cost metric can mirror reading times attested in the self-paced reading experiments of Crawley et al. (1990) who reported the reading time of the last sentence of the experimental items. A significant difference was reported between the ambiguous and the unambiguous condition in an overall variance analysis of the data.[12] The model also shows this difference. When we effected an analysis of variance on a by-item basis, the factor of ambiguity was highly significant (F = 299.5, df = 1, 39, $p < .001$). In Figure 2 the predictions of the model and the actual experimental reading times are plotted against each other.

Crawley et al. (1990) also compared reading times between the subject and the object assignment in the ambiguous and the unambiguous condition. They found faster reading times for subject assignment in the ambiguous condition, but this effect only showed in an analysis by participants and not by items (F_1 = 8.52, df = 1,47, $p > 0.1$; $F_2 < 1$). They did not find significant effects in the unambiguous condition, nor in the analysis by participants, nor in the analysis by items (F_1 = 1.55, df = 1,47, $p > 0.5$; F_2 = 1.08, df = 1,39, $p > 0.5$). Like Crawley et al. (1990), our model also showed a significant difference between subject and object

[9] We make the strong assumption that recovering the coreference chains in the psycholinguistic items is rather easy and does not cause much processing cost.

[10] We do not consider mentions having scores < 0.5, because it would mean that mentions that are classified 'negative' (probability less than 50%) could be of much as an influence as candidates being classed positive. We consider that negatively classified mentions do not add much to the competition there is between antecedent candidates.

[11] Note that the pronoun cannot be expletive (*i.e.* non-referential), because this type of pronoun is not annotated as a mention in the corpus and thus not considered by the system.

[12] We do not report the F-statistic here, because only the statistics for a by-subject analysis were reported.

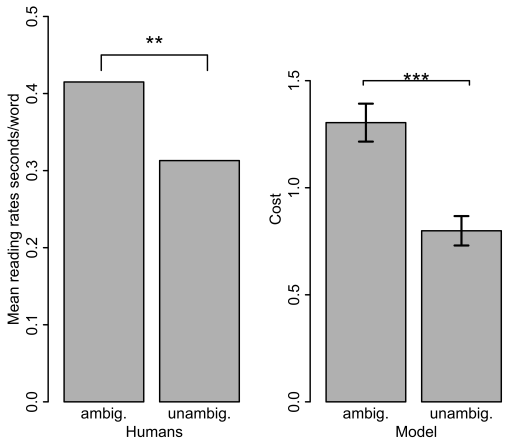


Figure 2: The model's prediction of processing cost against the reading times per word recorded by Crawley et al. (1990) for the ambiguous and the unambiguous condition of experiment 1. For the cost predicted by the model 95% confidence intervals are given. Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.

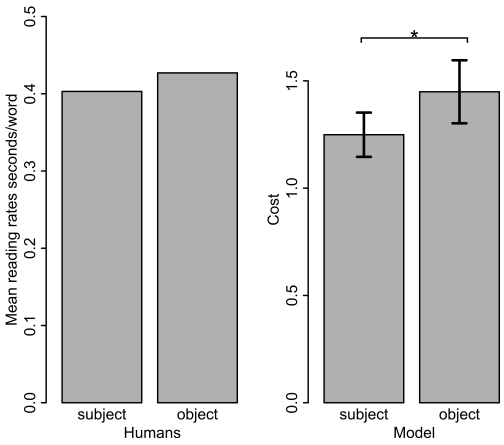
assignment in the ambiguous condition ($F = 4.23$, $df = 1, 38$, $p < .05$), but in an by-item analysis. For the unambiguous condition however, our results do not match Crawley et al. (1990)'s: we found a highly significant effect for the by-item analysis[13] ($F = 24.43$, $df = 1, 33$, $p < .001$). In Figure 3 the results for the subject and object assignment are plotted.

5.3 Discussion

Our cost metric is capable of mirroring the reading times of ambiguous versus unambiguous items and the reading times of items with subject and object antecedents in the ambiguous condition. However, in the unambiguous condition we found an effect that was not observed in the human data. We think that this can be explained by the strength of the gender and number features in our system. As the automatic gender and number feature assignment introduced some noise in our data, we think our model estimated these features lower than they should be, preventing them from erasing the influence of the Parallel Function Strategy and the Subject Assignment Strategy.

[13] In this analysis, items for which the resolver responded *None* were treated as missing values.

(a) Ambiguous Condition



(b) Unambiguous Condition

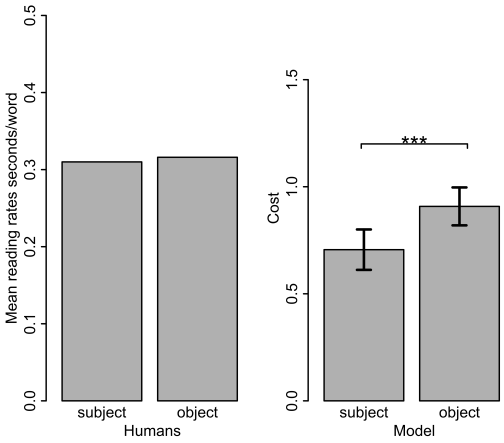


Figure 3: The model's prediction of processing cost against the reading times per word recorded by Crawley et al. (1990) for subject and object assignment in the ambiguous and the unambiguous condition of experiment 1. For the cost predicted by the model 95% confidence intervals are given.

6 General Discussion

The contribution of our model is its ability to quantify the strength of the factors of influence and its simple architecture that allows to incorporate easily new factors. The model also has the potential to explain human processing cost, because we were able to formulate a metric based on it that mirrored reading times recorded in the experiments of Crawley et al. (1990). Our results confirmed our idea that the competition between antecedent candidates can cause processing cost.

Our model can help in the psycholinguistic community to clarify statements about the exact nature of the involved factors. Indeed, when doing

the implementation of the model, many questions about how the features should be implemented arose. For example, implementing the parallel function turned out to be less straightforward than initially expected. We had to choose if we implemented it as a binary feature (the parallel function can only operate if the syntactic paths of both mentions are exactly the same), or a continuous feature (the similarity between the syntactic functions of the two mentions is what is relevant). Choices of this kind are very important when the modeling is done and inevitable. Of course, they are also relevant at the time of the design of the experimental items, but they can be overlooked more easily. The model also points out that in spite of the efforts of the experimenters to keep the items in one condition as similar as possible, many factors not included in the experimental design can still have an influence on the computational model and likely on the human participants as well. Let's take for example the items of Crawley et al. (1990): some items used proper nouns for the characters, whereas others contained only definite descriptions. This is likely to have an influence on the experienced difficulty, as suggested by the weights in Table 1, but also by theories such as the Accessibility Theory (Ariel, 1991) that states that different kinds of referential expressions are more or less accessible in memory for pronoun resolution. By detecting such things, we show that computational models can be a complementary means for psycholinguistic research.

As a future direction for our work, we plan to enhance our model, so that it would give a probability distribution over antecedent candidates in a more direct way. For the moment, as explained in section 5.1, we have to forge scores outputted by the resolver into a probability distribution, but it would be more elegant if this distribution came directly from the resolver.

We also plan to investigate further the way we define the cost metric. The idea to use entropy as a measure of uncertainty, or competition, is inspired by cost metrics for syntactic structure based on probability distribution, such as surprisal theory (Hale, 2001; Levy, 2008), predicting higher cost for unexpected syntactic structures, or the entropy reduction hypothesis (Hale, 2003; Hale, 2006), giving high cost at points where a lot of disambiguation is done. For the moment we only applied the notion of entropy, but we want to investigate if a notion of surprisal is applicable as well.

Finally, we plan to extend our model to other types of preferences. We would like for example to integrate discourse relations — that have been shown to have a great influence on pronoun resolution (Kehler and Rohde, 2013) — into our model. An even bigger challenge is to also integrate semantic information into the model. Another type of extension of our model is to get out of the experimental items and test our model on corpus data. We plan for example to test if our model can contribute to explain word by word reading times recorded on corpus — such as the Dundee eye-tracking corpus (Kennedy et al., 2003) — by adding it as a factor to a model including other factors explaining reading time, such as surprisal and word length.

7 Conclusion

In this paper we showed how a computational model can mirror human preferences in pronoun resolution and reading times with a cost metric based on the notion of entropy. We can see that the weights of the features learned on corpus correspond quite accurately to the influence of preferences in human pronoun resolution. We argue that our model will also be able to mirror other human preferences, provided we can learn the adequate features on corpus. A direction of future work is to enhance our multifactor model by more of these kinds of preferences, so that it will account for more and more preferences in pronoun resolution. We plan to ultimately test this model on reading times recorded on corpus.

Acknowledgments

We thank the three anonymous reviewers of the EACL Student Research Workshop, as well as or colleagues Tal Linzen, Maximin Coavoux and Sacha Beniamine for their comments, questions and suggestions on the paper. We also thank Adeline Nazarenko, our thesis co-director, her lab — the *Laboratoire d'Informatique de Paris Nord* — and the members of our thesis advisory committee — Saveria Colonna and Isabelle Tellier — for their support on this work. Finally, we thank our lab engineer Doriane Gras, for her help with the statistics. This work was supported by the *Labex Emperical Foundations of Linguistics* (ANR-10-LABX-0083) and the doctoral school *Frontières du Vivant*.

References

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5):443–463.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 33–40. Association for Computational Linguistics.

Rosalind A. Crawley, Rosemary J. Stevenson, and David Kleinman. 1990. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 19(4):245–264.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Pascal Denis and Jason Baldridge. 2007. A ranking approach to pronoun resolution. In *IJCAI*, pages 1588–1593.

Amit Dubey, Frank Keller, and Patrick Sturt. 2013. Probabilistic modeling of discourse-aware sentence processing. *Topics in cognitive science*, 5(3):425–451.

Stefan L. Frank, Mathieu Koppen, Leo G. M. Noordman, and Wietske Vonk. 2007. Coherence-driven resolution of referential ambiguity: A computational model. *Memory & cognition*, 35(6):1307–1322.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

Jerry R. Hobbs. 1976. Pronoun resolution. research report76-1. new york: Department of computer science. *City University of New York*.

Andrew Kehler and Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67. Association for Computational Linguistics.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 135–143. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2009. A deeper look into features for coreference resolution. In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 29–42. Springer.

Claude E. Shannon and Warren Weaver. 1949. *The mathematical theory of communication*. University of Illinois Press.

Amy Sheldon. 1974. The role of parallel function in the acquisition of relative clauses in english. *Journal of verbal learning and verbal behavior*, 13(3):272–281.

Ron Smyth. 1994. Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, 23(3):197–229.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Rosemary J. Stevenson, Alexander WR Nelson, and Keith Stenning. 1995. The role of parallelism in strategies of pronoun comprehension. *Language and Speech*, 38(4):393–418.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. An np-cluster based approach to coreference resolution. In *Proceedings of the 20th international conference on Computational Linguistics*, page 226. Association for Computational Linguistics.

A All features that were considered before model selection

Feature	Decision
match in gender	keep
match in number	keep
m_1 is a subject	keep
match in syntactic path	keep
m_1 is a common noun	keep
m_1 is a proper name	keep
m_1 is a possessive pronoun	keep
m_1 is a personal pronoun	keep
mentions between m_1 and m_2	not significant
words between m_1 and m_2	keep
m_1 & m_2 in the same sentence	keep
length of syntactic path m_1	keep
m_1 is determined	keep
m_1 is undetermined	keep
m_1 has a demonstrative determiner	keep
m_1 spans m_2	keep
length of words of m_1	keep
number of occurrences of m_1 in the text	not significant
m_1 is a location	not significant
m_1 is a work of art	not enough data
m_1 is a geopolitical entity	keep
m_1 is an organization	not enough data
m_1 is a date	keep
m_1 is a product	not enough data
m_1 is a NORP[14]	not enough data
m_1 is a language	not enough data
m_1 is money	not enough data
m_1 is a person	not significant
m_1 is a law	not enough data
m_1 is an event	not enough data
m_1 is a quantity	not enough data

B Gender and Number Annotation

We used the Bergsma and Lin (2006) gender information, that provides counts of word forms occurring as respectively male, female and neuter gender on the web, to annotate the mentions in our corpus. More precisely, we took the three lists of unigrams (one for each gender) from the Stanford Core NLP Toolkit (Manning et al., 2014) that was compiled from the Bergsma and Lin (2006) gender information to annotate each token of a mention in our corpus with gender if it occurred in one of the lists. Then we propagated the gender of the head token to the entire mention. Finding the head of a mention was done using a heuristic: the head is the last word of the mention, except if there is a prepositional phrase inside the mention, in the latter case the head of the mention is the word before any prepositional phrase.

The number annotation was only done for tokens that were common nouns and proper names. In the tag set of the corpus singular common nouns are tagged as *NN*, singular proper names as *NNP*, plural common nouns as *NNS* and plural proper names as *NNPS*. We used these tags to assign number to tokens. Then, we proceeded with the same *head heuristic* as for the gender feature to assign number to the entire mention.

[14] nationalities, organizations, religions, and political parties