name is, on average, higher in test than in training. The results plotted in Figure 1 show that it is not a question of which algorithm is better, but rather that there are different cases where one approach is preferred over the other. The problem we face is deciding when it is appropriate to use one or the other.

To induce from the corpus itself when a piece of context is or is not a relevant context requires deep ontological inferences and a very powerful tool of semantic analysis of the context. Consider for example two words denoting profession, "doctor" and "researcher", and their possible modifiers "internist", "neurosurgeon", and "professor" and "PhD". In the first case it is certain that the coreference is not possible, while in the second the coreference is very probable. To find out such relationships is computationally very hard. However, the analysis carried out further shows that we can avoid making such computations in most of the cases.

The number of different persons is a parameter that cannot be known beforehand. However, not all the names behave alike with respect to coreference. There are noticeable differences between names; for example less than 5 000 first names cover approximately 96% of the total of first names, while for the same percentage of coverage more than 70 000 of last names must be considered (Popescu et al. 2007). Let us call perplexity of a name the number of different persons that carry it. The search space depends directly on the name perplexity. The bigger the perplexity, the larger the amount of information required for the correct coreference must be. It seems natural that the amount of contextual evidence required by a PCDC depends on the name perplexity.

In order to evaluate the relationships between the context and the name perplexity, we need an annotated corpus. We have used the I-CAB corpus (Magnini et al. 2006), which is a four-day news corpus fully annotated, coreference relationships included. The documents in this corpus are entire pieces of news. For each PNM we have counted how many contexts containing specific information about the person carrying the respective name is present in that particular document. There are many types of contexts that refer to a person, but some of these types are very infrequent. We considered only those types of information that are present at least 5% of the times in the context surrounding a PNM. Table 2 presents the results of this investigation.

	occ	diff occ	entities
First Names	2299	676	1592
Last Names	4173	1906	2191
Middle Name	116	44	41
Activity	973	322	569
Affiliation	566	399	409
Role	531	211	317
Family Relation	133	40	94

Table 2. Name perplexity and context

On the second column the total number of occurrences is listed, on the third column how many of these occurrences have different values (no case sensitive string match), and on the fourth column the number of different persons (Entities) having that information. The entries "activity", "affiliation", and "role" represent pieces of context where the respective information is directly expressed (no inferences). We call this type of context professional context and for approximately 30% of the PNMs, one of the above three types of professional contexts is present.

The perplexity of the first names, computed as the ratio between the fourth column and the third column is two times bigger than the perplexity of the last names. The lowest name perplexity is obtained by the names having a middle name - a name with at least three tokens – and it is very close to 1 (1.07). Comparatively, the highest perplexity of two tokens name is 3. The relationship between the number of tokens of a name and its perplexity is straightforward: for names with more than four tokens the perplexity is 1 in 99,6% of the cases (the name by itself is a relevant context for coreference).

In approximately 74% of the cases there is just one entity corresponding to a two-token name. Considering any two PNMs of the same name the similarity of two of the professional contexts guarantees the correct coreference. However, two professional contexts are present in only 4% of the cases. There are just four cases when considering just one professional attribute was misleading, and all these cases are high perplexity names. Moreover, in the case of many low perplexity names, the contexts could be minimally similar in order to correctly corefer any two PNMs of that respective name.

This analysis shows that there is a direct relationship between the name perplexity and the relevant coreference context. However, the average figures are not very informative, as the variance of perplexity is very high. Rather than fo-