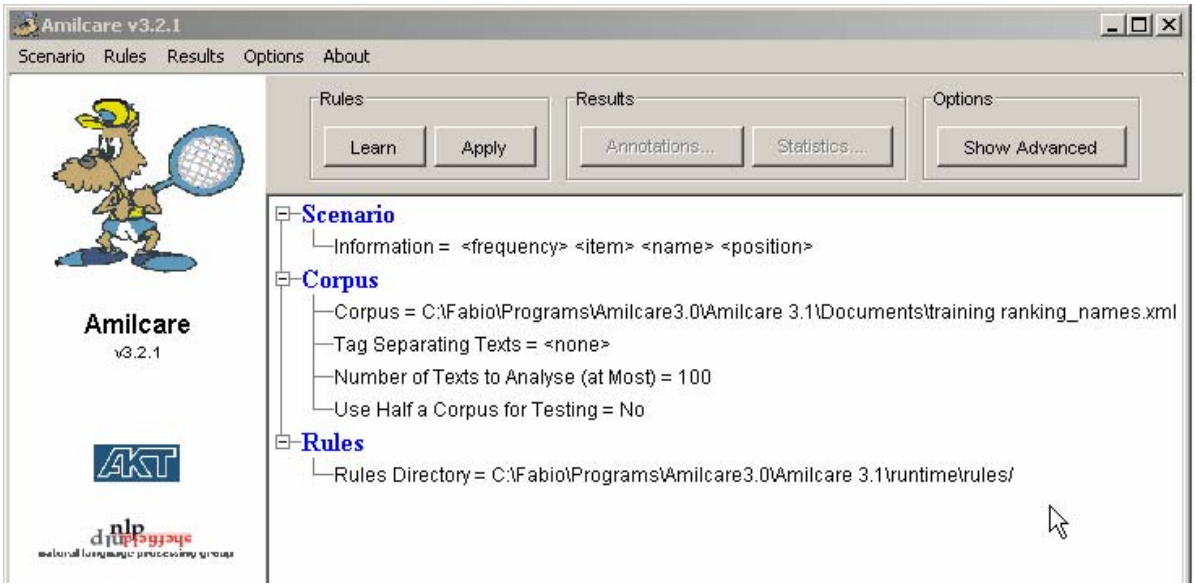
Tuning Amilcare

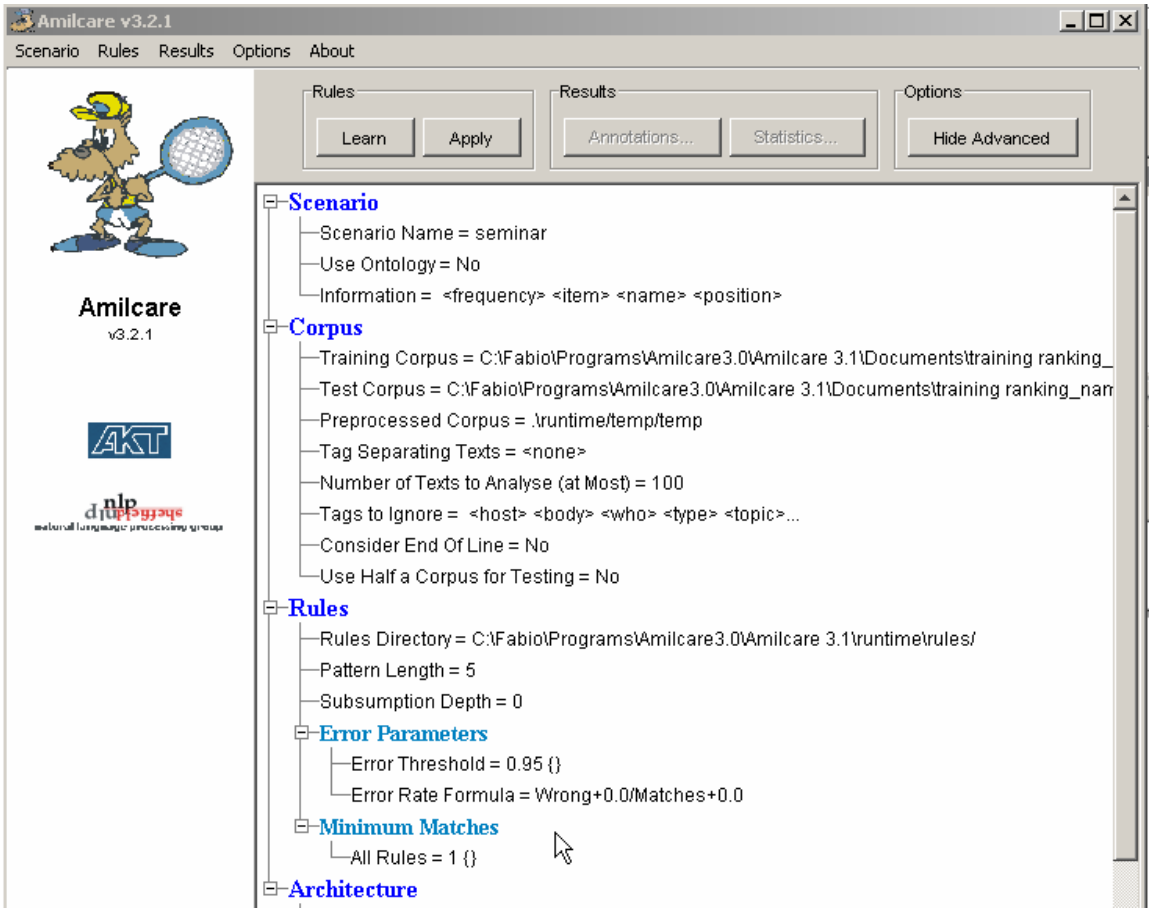
Running Amilcare without accuracy tuning can lead to **some very disappointing results.** You will need to tune the accuracy thresholds for optimizing results.

If you need to tune the thresholds here are the instructions. It may seem daunting at first, but it generally requires just two or three runs and it becomes very easy after a while.

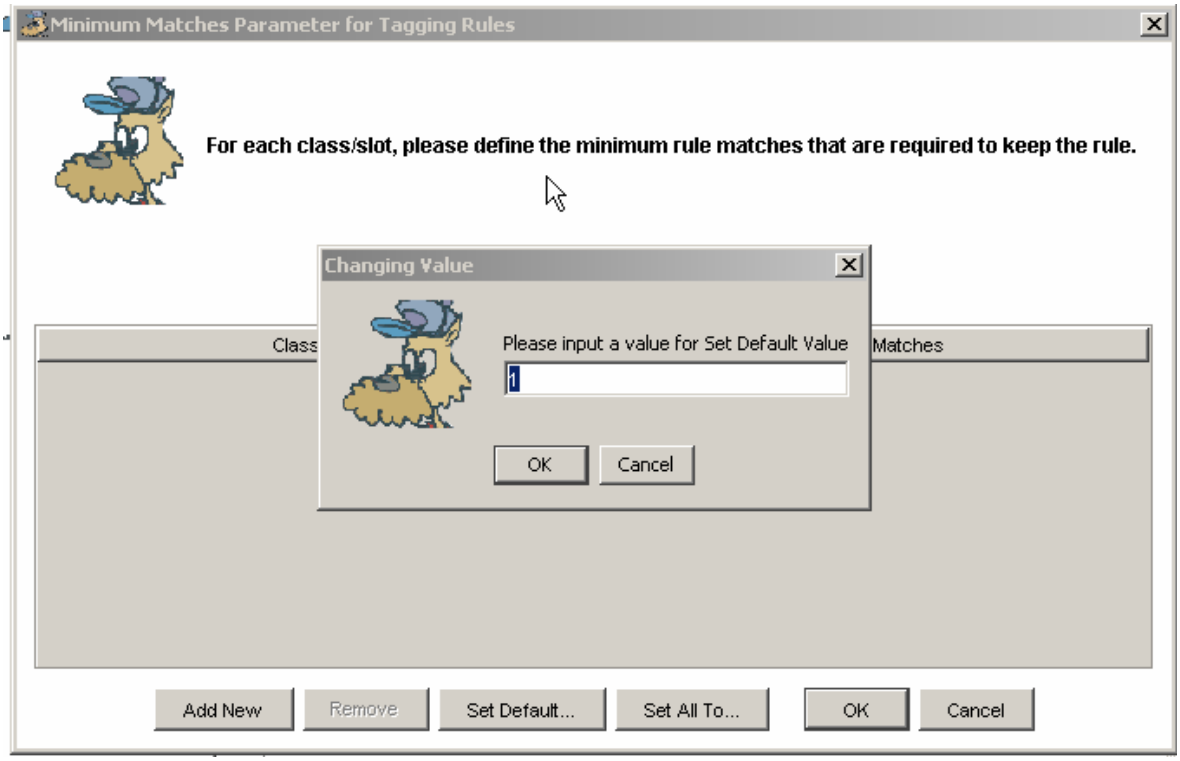
Open Amilcare.



Click the button show advanced on the top right corner. The window changes as follows:



set Minimum Matches to 1 (the default is 3 and will return not many results with that one). Click on minimum matches, click on “Set Default”, the following window will open:

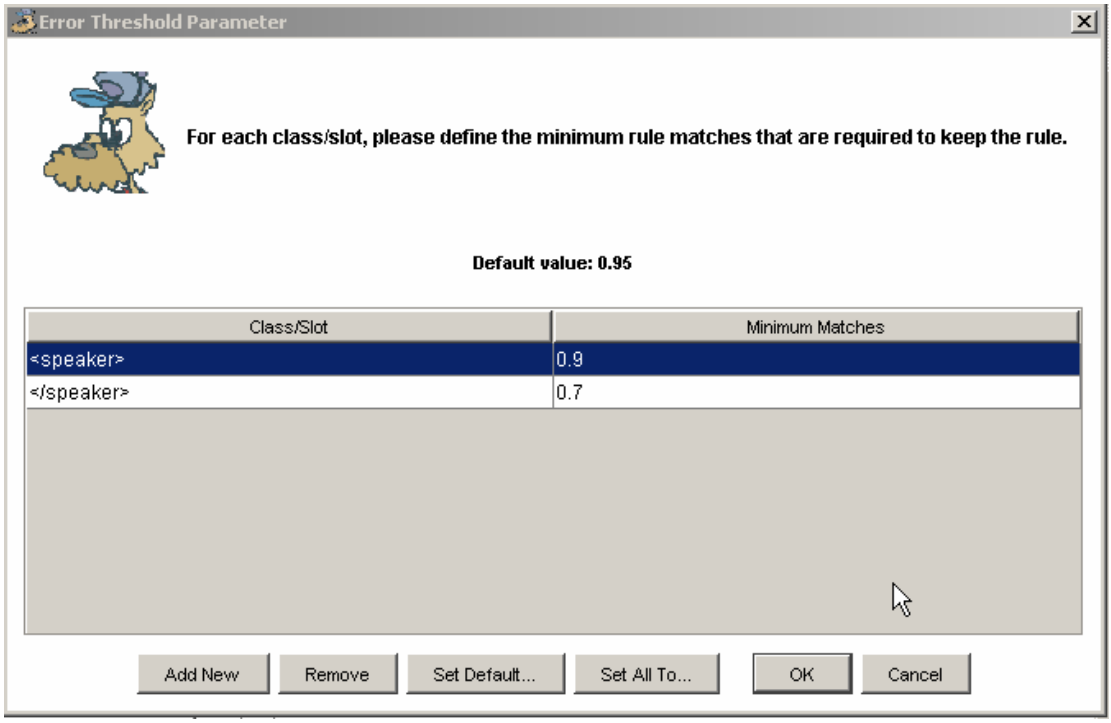


Then you need to tune the accuracy thresholds.

Each tag threshold can be separately tuned; start tags (e.g. <speaker>) can have different thresholds from their corresponding end tag (e.g. </speaker>).

You will need a **training corpus and a tuning corpus** used to tune the thresholds (called test corpus hereafter). Train and test sets must not have any intersection.

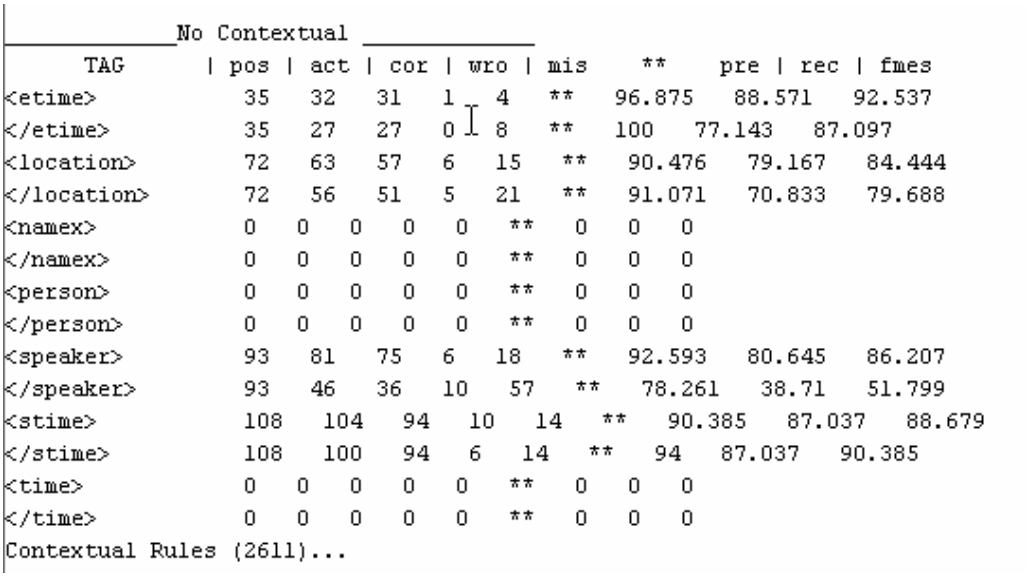
To set the thresholds you need to click on “error parameters” in the main window and the following will appear (with thresholds for <speaker> and </speaker> already inserted in this case).
NOTE! When you insert a threshold, always be sure to terminate your input with the <return> key, otherwise the value could be missing!!!



Start with all thresholds set to 1.0 for start tags and 0.9 for end tags.

Tuning

- Train Amilcare and test it on the sample test corpus
- Look at the intermediate results on the test corpus. I mean have a look at the trace Amilcare produces. Look for tags that report low recall before contextual rule applications (first table in the trace, title is “no contextual” – see figure below) and reduce the error threshold for those tags (e.g. to 0.8). For example you will for sure noticing that results tend to be better for start tags (e.g. <speaker>) and worse for end tags (</speaker>). You generally get lower recall for end tags. In this case you need to make the error threshold smaller (e.g. if it is 1.0 put it to 0.8). You must do that for every tag you have low recall for (not only end tags, in case). For example in the table below, </speaker> has very low recall (38.71). This means that the threshold should be reduced.



No Contextual									
TAG	pos	act	cor	wro	mis	**	pre	rec	fmes
<etime>	35	32	31	1	4	**	96.875	88.571	92.537
</etime>	35	27	27	0	8	**	100	77.143	87.097
<location>	72	63	57	6	15	**	90.476	79.167	84.444
</location>	72	56	51	5	21	**	91.071	70.833	79.688
<namex>	0	0	0	0	0	**	0	0	0
</namex>	0	0	0	0	0	**	0	0	0
<person>	0	0	0	0	0	**	0	0	0
</person>	0	0	0	0	0	**	0	0	0
<speaker>	93	81	75	6	18	**	92.593	80.645	86.207
</speaker>	93	46	36	10	57	**	78.261	38.71	51.799
<stime>	108	104	94	10	14	**	90.385	87.037	88.679
</stime>	108	100	94	6	14	**	94	87.037	90.385
<time>	0	0	0	0	0	**	0	0	0
</time>	0	0	0	0	0	**	0	0	0

Contextual Rules (2611)...

Then train again and test. Recall should in principle increase if not rocket. Higher recall for single tags (start or end) should affect both precision and recall. In my experience you have a larger gain in recall and a mixed effect on precision. It can increase as well or reduce a bit, but never dramatically collapse.

- In order to check if you need more training, look at the rule coverage after training in Amilcare. If they tend to cover only a very limited number of cases (say 1-3% of the training corpus) and you do not have a consistent set covering some 10% or more (say 3-5 rules), I think there is a problem with the corpus (data sparsity). Otherwise it is ok.
- If you still get low results on the test, it may be because the training corpus is not representative: how you select training and corpus should be completely random so to allow a reasonable distribution of cases. Pseudo random selections like separating the corpus alphabetically by the file name (say a-m for training and n-z for test) or via date of creation could take in principle to disastrous results. This is because the training or the test can be influenced by local problems you do not have control on and that can influence the corpus badly. For example all the files with name xxx-1 xxx-2 xxx-3, etc. could be similar and maybe introduce unbalanced features. I can tell you a story: when an Italian financial news wire company gave us to train on their corpus produced by three consecutive days of feed we thought it was ok. We split the corpus in train and test by separating them randomly. We train and tested the system and the results were beyond our expectations. When the system went online on a final test it gave horrible results. We inspected the documents and we discovered that the documents they had sent us were influenced by the fact that in those three days there were (1) an attempted hostile takeover of the Italian Telecom (5th most important telcom company in the world) and (2) a major tax deadline for Italy, so most of the documents were about that two topics. Consistency within the corpus: high; generelaty of induced rules: nil. They had annotated the documents so we had not paid enough attention. This story does not probably apply to your case, but I think it can shed light on problems you can have in choosing test/training corpus.

Try and do a test. If you need help, send me Amilcare's trace for both training and test and I will tell you what to do.