1. Give some subfields of linguistics where only corpus analysis is possible, and not introspection. Explain why it is so.

2. The notion of representativity is an important notion of corpus linguistics. Explain quickly but clearly:
   – what is the aim of getting representativity;
   – what are the ways to ensure representativity;
   – an example where the lack of representativity could give misleading results.

3. The analysis of a corpus of 100 sentences (Le Monde) indicates that the average length of the sentences is $\bar{x}_0 = 29.24$, with a variance $s_0^2 = 238.7$.
   – What are the main properties of the normal distribution?
   – What are the min and max values for the mean of the population $\mu$ so that $\bar{x}_0$ is in the 90% proportion (that is the 90% confidence limits for the mean)? Justify your computation.
   – The analysis of another corpus (of size 100 too) indicates $\bar{x}_1 = 20.39$ words per sentence, with a variance of $s_1^2 = 87.84$. Can these two samples come from a random sample in the same population? Justify your answer (choice of the test, null hypothesis, significance level…).

4. Modalities for three different genres (fiction, adventure and romance) of the Brown Corpus are distributed along the mood (indicative and conditional) as in table 1.
   (a) Does the usage of the mood for the modalities vary along the genres? With which significance level?
   (b) If you consider only romance and mystery, do you have the same conclusions with a significance level of 10%? Of 5%?

**Reminder**

Standard error of the difference of the means: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
<table>
<thead>
<tr>
<th></th>
<th>indicatif</th>
<th>conditionnel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>adventure</strong></td>
<td>135</td>
<td>422</td>
</tr>
<tr>
<td><strong>romance</strong></td>
<td>176</td>
<td>534</td>
</tr>
<tr>
<td><strong>mystery</strong></td>
<td>106</td>
<td>418</td>
</tr>
</tbody>
</table>

**Tab. 1 – Distribution of the modalities (Brown Corpus)**