Sujet d'examen : Spécialité : Traitement automatique des Langues
I-TAL-3.7 : Corpus linguistics, linguistic resources and normalisation

For every question, we recommend **to explain** and **to precisely justify** all the reasonings that lead you to the answer. The quality, the clarity and the precision of these elements will be taken into account (remember that clarity and concision often go hand in hand).

Don't forget to read all the examination questions before starting to answer.

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 μ 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: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

	indicatif	conditionnel
adventure	135	422
romance	176	534
mystery	106	418

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