Sujet d'examen : Spécialité : Traitement automatique des Langues
I-TAL-3.11 : 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 a quick answer to the following questions:
   - The Penn treebank has been annotated with different kinds of annotations. Give three of them, with examples (of course you can use artificial data and tags in your examples) and the possible outcomes of such annotations.

     POS annotation, syntactic annotation, semantoic role annotation and discourse units/connective annotations. See the slides for examples. Possible outcomes: use syntax to desambiguate semantic roles or identify discourse units.

   - What are the differences between intensional and extensional lexicons.

     *An intensional lexicon gives a normalized form (a lemma) for each lexical item, then a description of **how** to build the flexed from from it. An extensional lexicon contains all the flexed forms.*

2. What is the difference between competence and performance? How does it translate in corpora? How is that distinction taken into account in corpus linguistics?

   The competence is an idealized capacity of analyzing the language, whereas the performance is the way this knowledge is applied and expressed. The latter is observable and can be modified by the environment (social, physical...). Corpora only have access to performance.
   Elaboration on:
   - *representativity of corpora*
   - *in some area, introspection is unavailable (language acquisition by children, dead languages, phonetics...)*
   - *the systematicity makes it in some sense more reliable than introspection*
   - *there are ways to generalize data from a corpus in order to make infered structures applicable to new items (learning)*
   - *large data are available*

3. A study is run on the length of french words. The average length of 100 words of a newspaper is $\bar{x}_0 = 4.72$ letters (standard deviation is $s_0 = 2.85$).
   - Let us assume that μ, the mean of the population is $\mu = 4.90$. Between which values (μ being right in the middle) 95% of samples of size 100 have their average? Justify your answer (choice of the score, choice of the values, possible assumptions). Does $\bar{x}_0$ belong to this interval?

     With samples of this size (≥ 30), the means of samples follow the normal distribution, hence we can use a z-score. We will approximate the standard error of the mean by $s_e = \frac{s_0}{\sqrt{100}}$. To get the 95% we need to have $z = \frac{\bar{x}-\mu}{s_e}$ between $-z_0$ and z_0 with $z_0 = 1.96$ (that is 0.4750 on the left and on the right). So we have:

$$\begin{array}{lccccc} & -z_0 & \leq & \frac{\bar{x}-\mu}{s_0} * 10 & \leq z_0 & \\ \Leftrightarrow & \mu - \frac{z_0 * s_0}{10} & \leq & \bar{x} & \leq & \mu + \frac{z_0 * s_0}{10} \\ \Leftrightarrow & \mu - \frac{1.96 * 2.85}{10} & \leq & \bar{x} & \leq & \mu + \frac{1.96 * 2.85}{10} \\ \Leftrightarrow & \mu - 0.56 & \leq & \bar{x} & \leq & \mu + 0.56 \\ \Leftrightarrow & 4.16 & \leq & \bar{x} & \leq & 5.28 \end{array}$$

   - We don't make any assumption on μ anymore. What is the lowest (μ_1) and the highest (μ_2) values of μ for which $\bar{x}_0$ would be in the 95% proportion? (They are called the 95% confidence limits for the mean.)

     Now, μ is the unknown. μ_1 is such that $z_1 = \frac{\bar{x_0}-\mu_1}{s_e} = z_0$ and $z_2 = \frac{\bar{x}-\mu_2}{s_e} = -z_0$. That is:

$$\begin{array}{llll} \mu_1 & = \bar{x}_0 - z_0 * s_e & = \bar{x}_0 - \frac{z_0 * s_0}{10} & = \bar{x}_0 - 0.56 \quad = 4.16 \\ \mu_2 & = \bar{x}_0 + z_0 * s_e & = \bar{x}_0 + \frac{z_0 * s_0}{10} & = \bar{x}_0 + 0.56 \quad = 5.28 \end{array}$$

   - What are the 99% confidence limits (μ'_1 and μ'_2)?

     The same computatino as before holds, except that now we have $z'_0 = 2.58$ instead of $z_0 = 1.96$. Then:

$$\begin{array}{llll} \mu'_1 & = \bar{x}_0 - z'_0 * s_e & = \bar{x}_0 - \frac{z'_0 * s_0}{10} & = \bar{x}_0 - 0.74 \quad = 3.98 \\ \mu'_2 & = \bar{x}_0 + z'_0 * s_e & = \bar{x}_0 + \frac{z'_0 * s_0}{10} & = \bar{x}_0 + 0.74 \quad = 5.44 \end{array}$$

   - Assuming the same deviation, what is the required size for a sample **to get** μ_1 **and** μ_2 **again** for the 99% confidence limits for the mean?

     Let n be this size. Replacing 10 by $\sqrt{n}$ in the previous answers, we get (with a new standard error s'_e):

$$\begin{array}{lll} \mu_1 & = \bar{x}_0 - z'_0 * s'_e & = \bar{x}_0 - \frac{z'_0 * s_0}{\sqrt{n}} \\ \Leftrightarrow \frac{z'_0 * s_0}{\bar{x}_0 - \mu_1} & = \sqrt{n} & \\ \Leftrightarrow n = \frac{(z'_0 * s_0)^2}{(\bar{x}_0 - \mu_1)^2} & & = \frac{(2.58 * 2.85)^2}{(4.72 - 4.16)^2} = 172.4 \end{array}$$

– Another sample of size 100 taken from a tourist guide results in an average $\bar{x}_1 = 5.63$ ($s_1 = 3.26$). Can these two samples have been randomly made out of the same population? Justify you answer and describe your reasoning (choice of the test, null hypothesis, chosen significance level, etc.)

The length of words is an ratio variable, hence we can choose a parametric test to test the difference between two populations. The samples are large enough, and not correlated, hence the difference of the means follows a normal law with 0 *as mean and* $s_d = \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}$. *as standard deviation.*
The null hypothesis is the following: the two samples have been taken from the same population.
The z-score is $z = \frac{\bar{x_1} - \bar{x_0} - 0}{s_d} = 10 * \frac{\bar{x_1} - \bar{x_0}}{\sqrt{s_1^2 + s_0^2}} = 10 * \frac{1.09^2}{\sqrt{2.85^2 + 3.26^2}} = 2.74.$
We choose a one-tailed test, hence we can reject the null hypothesis with a 0.64% *level of confidence.*

4. If we look at the mood of modals with respect to their usage in positive and negative clauses, we get the results of table 1.
   (a) What can you say **for modals** about the usage of negation with respect to the mood of modals? With which level of significance?

Since we have nominal variables, to compare the two samples, we need a non-parametric test such as the χ^2 *test.*
The null hyopthesis is: the positive and negative modals are taken from the same population (no distinction between ind. or cond.)
We have the following obervations:

	ind.	*cond.*	*Total*
pos.	*465*	*371*	*836*
neg	*42*	*48*	*90*
Tot.	*507*	*419*	*926*

and the following expectations:

	ind.	*cond.*	*Total*
pos.	$\frac{507*836}{926} = 457.72$	$\frac{419*836}{926} = 378.28$	*836*
neg	$\frac{507*90}{926} = 49.28$	$\frac{419*90}{926} = 40.72$	*90*
Tot.	*507*	*419*	*926*

Hence $\chi^2 = \Sigma \frac{(O-E)^2}{E} = 2.63$.*.With 1 degree of freedom, it means we can not reject the null hypothesis with a confidence level less than about* 10%.

   (b) What can you say if you compare the negative and positive distribution with respect to

modals (with no mood distinction) and regular verbs?

With the same reasonning as before, we have the following observations:

	modals.	*regular.*	*Total*
pos.	*836*	*8325*	*9161*
neg	*90*	*293*	*383*
Tot.	*926*	*8618*	*9544*

and the following observations:

	modals.	*regular.*	*Total*
pos.	$\frac{926*9161}{9544} = 888.84$	$\frac{8618*9161}{9544} = 8272.16$	*9161*
neg	$\frac{926*383}{9544} = 37.16$	$\frac{8618*383}{9544} = 345.84$	*383*
Tot.	*926*	*8618*	*9544*

Now $\chi^2 = \Sigma\frac{(O-E)^2}{E} = 86.69$*..With 1 degree of freedom, it means we can reject the null hypothesis with a confidence level less than about almost* 0%*. (not even in the table).*

	modals		regular
	indicatif	conditionnel	
positive	465	371	8325
negative	42	48	293

TAB. 1 – *Distribution of positive or negative clauses (Penn treebank)*

Reminder

Standard error of the difference of the means: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$