Distribution of modal expressions of possibility and necessity in three encyclopedias covering two diachronic spans (18th and 21st centuries)

Corinne Rossari[*1], Jessica Chessex[1], Claudia Ricci[*1], Iveta Walther[1] and Dennis Wandel[1]

[1]University of Neuchâtel – Switzerland – corinne.rossari@unine.ch, jessica.chessex@unine.ch, claudia.ricci@unine.ch, iveta.walther@unine.ch, dennis.wandel@unine.ch

Abstract

The aim of this paper is to measure to what extent the modalities of possibility and necessity are represented in three French encyclopedic corpora covering two diachronic spans, the 18th and the 21st centuries, by applying a statistical approach to linguistic data. Using log-likelihood, we will measure over- and under-representation of morphological and lexical modal forms in those encyclopedias, as well as their association with other linguistic items. Correspondence analysis will then be used to give a more holistic view of these associations. It will be shown that the two types of modality studied are used differently in the 18th century encyclopedic corpus compared to the two corpora from the 21st century.

Keywords: modality, corpus linguistics, statistical approach, encyclopedias, diachrony

1. Introduction

We aim to use statistical methods on linguistic data to measure to what extent the modalities of possibility and necessity are represented in three major French encyclopedias: the *Encyclopædia Universalis* (Univ) and French *Wikipédia* (Wiki) for the 21st century, and the *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, also known as *Encyclopédie Diderot et D'Alembert* (DDA) for the 18th century. More precisely, we intend to identify to what degree these types of texts contain knowledge mediated by an assessment instance, as implied by the use of linguistic forms expressing possibility and necessity. With the help of quantitative analysis (Blumenthal, 2017; Charolles et al., 2017; Vigier, 2017), we intend to detect similarities and differences between these encyclopedias concerning their modes of presentation of knowledge. Our study will look at the canonical forms expressing possibility and necessity, namely the verbs *pouvoir* 'can', *devoir* 'must' and *falloir* 'must', but we will also explore non-canonical forms, such as the adjectives formed with the suffixes *-ible* and *-able*. Their meaning generally implies possibility (*faisable* is paraphrasable by 'can be done', *compréhensible* by 'can be understood') and, marginally, necessity (*condamnable* by 'must be condemned'). To evaluate the distribution of these modalities in each of the works considered, we will first measure over- and under-representation of these forms in the encyclopedias by calculating their specificity scores using the association measure of log-likelihood (LL) (Evert 2008), cf. sect. 1. In a second step, we will further refine the statistical methods. We will first calculate their association with other linguistics items by means of log-likelihood (sect. 2.1. and 2.2.). Then, we will use correspondence analysis (CA) (Desagulier, 2017; Greenacre, 2017) cf. sect. 2.3., to get a more holistic view of these associations. Crossing these statistical tools will give us the possibility to assess the variation in the use of modality in the three different encyclopedias.

[*] Speakers.

2. Over- and under-representation of modal forms in the three encyclopedias

To assess the use of the modalities of necessity and possibility in the three encyclopedias mentioned above, we have taken into account two different parts of speech: suffixes (*-ible* and *-able*) and verbal lexical forms (*devoir*, *falloir*, *pouvoir*). Both these categories can convey a meaning of possibility or of necessity. Cross-referencing of the data allows to pinpoint convergences between these two types of forms in each encyclopedia.

Figure 1:

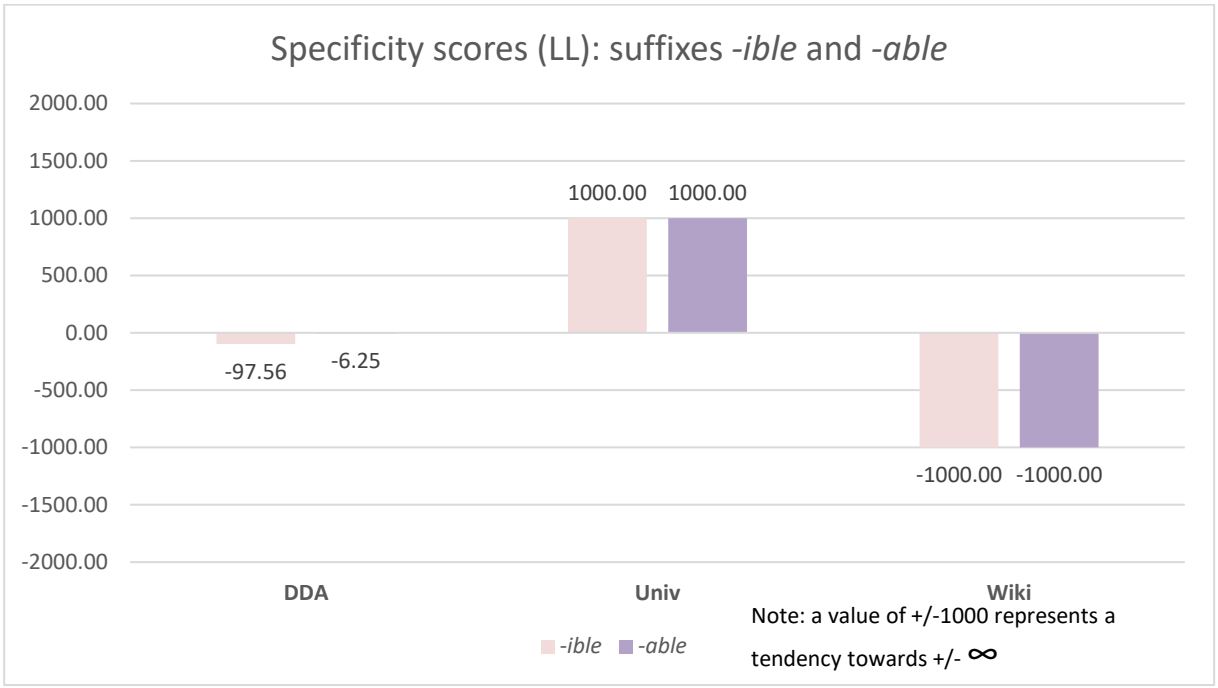
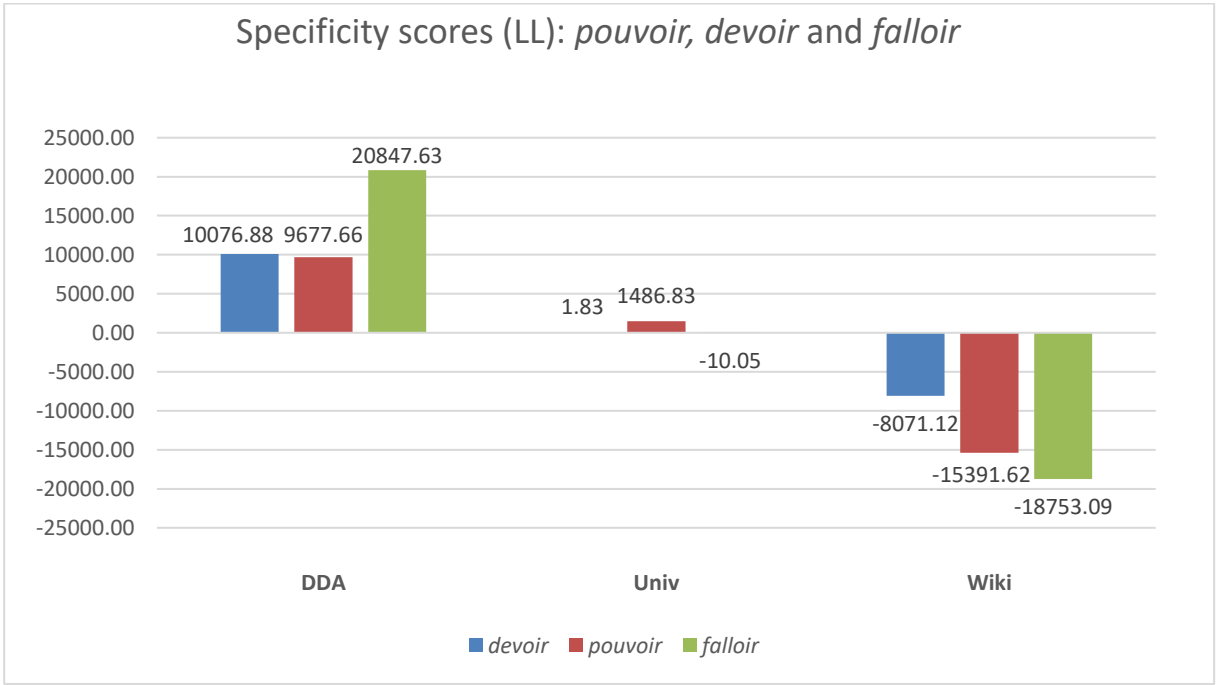


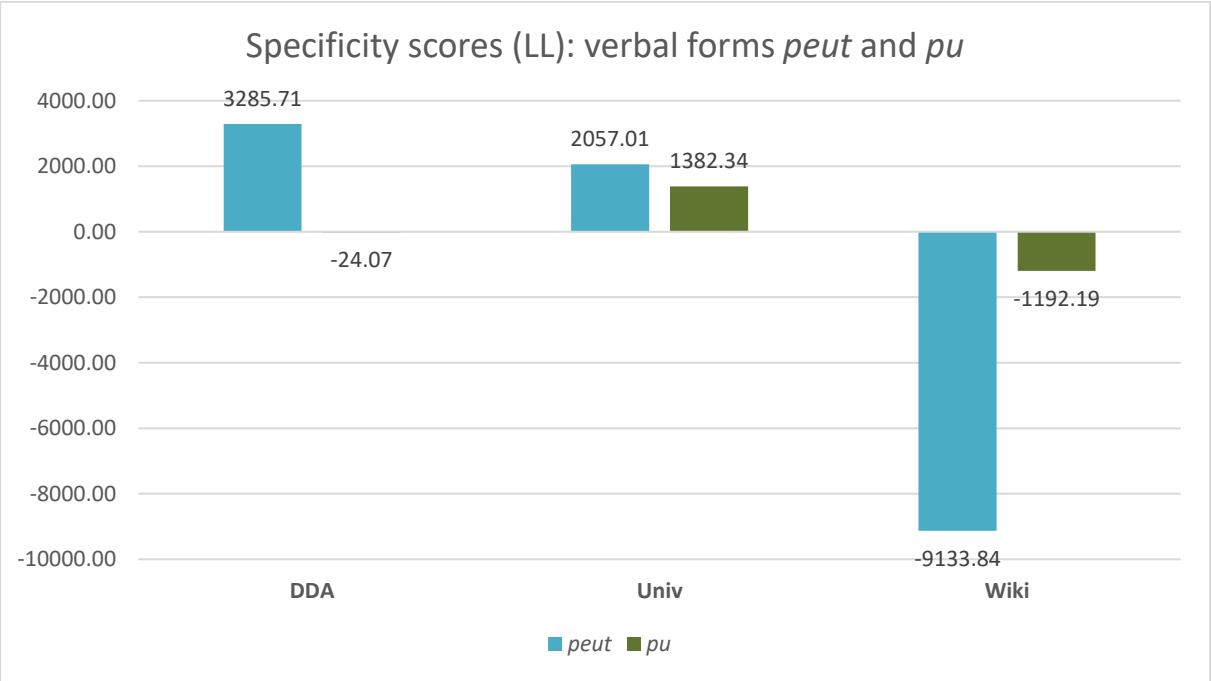
Figure 2:



First, we can observe that *-ible* and *-able* forms have homogeneous specificity scores: this indicates that the modal semantic feature that they share, regardless of the basis with which the suffixes are combined, plays a significant role in their use. Second, we can notice that the trends shown by the modal lexical forms are mostly similar to those shown by the *-ible* and *-able* suffixes, particularly in *Wikipédia*, where both forms are strongly under-represented. This similarity confirms that the modal semantic feature shared by these forms is essential for defining their use. Third, we notice that (i) in *Universalis*, the modality of possibility conveyed by the suffixes *-ible* and *-able* (in most of their uses) and by the verbal lexical form *pouvoir* is over-represented; (ii) both possibility and necessity conveyed by verbal lexical forms are over-represented in *Diderot et D'Alembert*.

Another comparison of specificity scores of some particular forms of the verb *pouvoir* (the indicative present, 3rd person singular *peut* and the past participle *pu*) seems to confirm the over-representation of the modality of possibility in *Universalis* (figure 3).

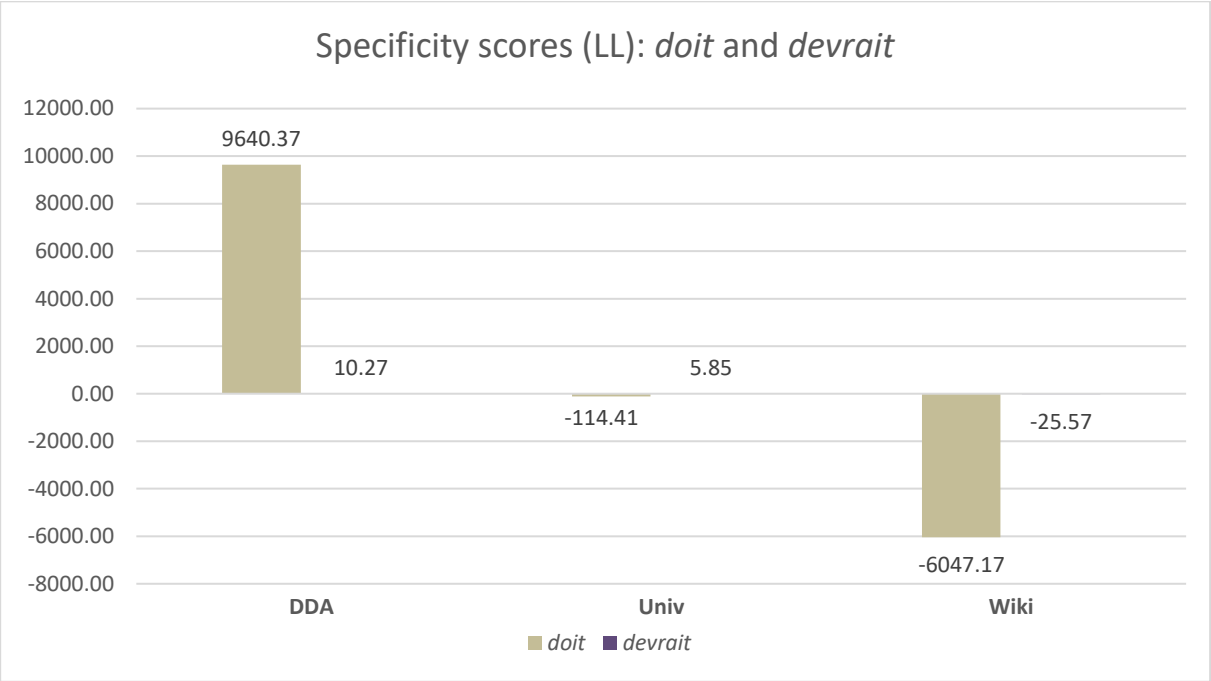
Figure 3:



First, showing a strong homogeneity in their specificity scores (which is consistent with the other forms tested), the measure verifies that the tendencies observed before are indeed linked to the semantic indications of the lemma. Second, once again, these forms are over-represented in *Universalis* compared to *Wikipédia* and *Diderot et D'Alembert*: in the latter, only *peut* is over-represented.

By extending our research to verbal forms conveying necessity (the indicative present, 3rd person singular *doit* and the present conditional, 3rd person singular *devrait*), we see that the 18th-century encyclopedia – which shows a predilection for the use of *devoir* in figure 2 – clearly stands out with values showing an over-representation for *doit* and a stronger LL value for *devrait* compared to the other two encyclopedias.

Figure 4:



To sum up, the specificity scores reveal tendencies particularizing some encyclopedias: (i) *Wikipédia* shows reticence to using both modalities (necessity or possibility); (ii) *Universalis* shows a preference for the modality of possibility; (iii) *Diderot et D'Alembert* shows a preference for the modality of necessity.

To further explore the use of modality in each encyclopedia, we use log-likelihood to identify linguistic items to which the modal lexical forms and the modal morphological suffixes are significantly attracted.

3. Measuring the association of modality with other forms

3.1 LL measures for lexical modality

Our methodology for this section is corpus driven. Among the collocates, we have selected the first 25 verbs according to their rank assigned by LL in each encyclopedia. We have then selected only verbs conveying an assessment of a state of affairs, in the sense of mediating a content, which we call 'cognitive verbs' referring to Rabatel (2003), who uses this label to designate verbs expressing a mental process. We have divided them into two categories: *dicendi* verbs (category 1) and thought activity verbs (category 2). For the first category we have selected the following five verbs: *dire* 'say', *citer* 'quote', *demander* 'ask', *parler* 'talk', *répondre* 'answer'; for the second category, eight verbs have been chosen: *juger* 'judge', *douter* 'doubt', *considérer* 'consider', *penser* 'think', *définir* 'define', *expliquer* 'explain', *conclure* 'conclude', *noter* 'note'. To verify that they are used to mediate a content, we have manually sorted all the occurrences to keep only those where the verb has a *dicendi* meaning or indicates a thought process. For instance, we have excluded occurrences such as '*répondre* d'une action' or '*demander* sa part d'héritage' ('to be *held accountable* for an action' or '*claim* one's share of the inheritance').

Comparing the results from the three encyclopedias, we can observe tendencies that we could not deduce from the specificity scores presented above. On the one hand, there are disparities between the modality of necessity and the modality of possibility, revealing a large preference for the modality of possibility for both categories of verbs in all encyclopedias. On the other hand, the modality of necessity is associated with more *dicendi* verbs in *Universalis* than in *Diderot et D'Alembert*, whereas the specificity scores (see figures 2 and 4) show it as being over-represented in the latter.

Figure 5:

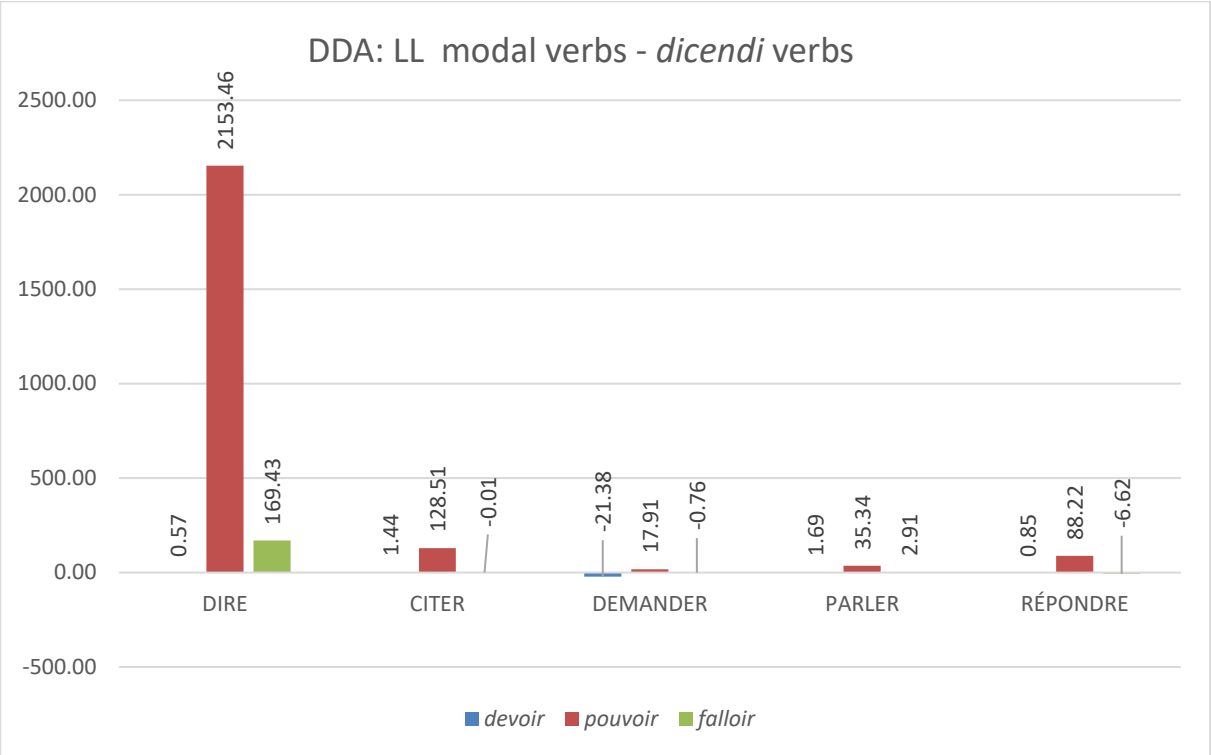
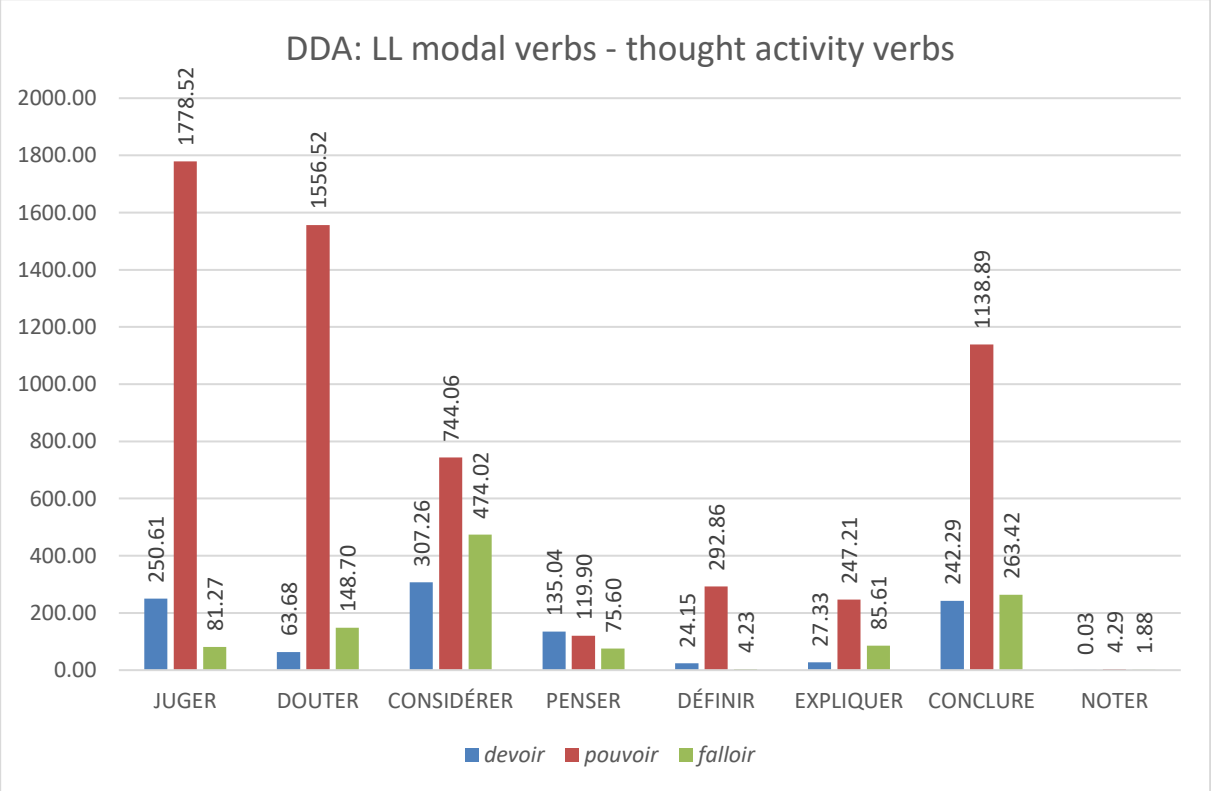


Figure 6:



Diderot et D'Alembert stands out concerning necessity verbs in association with *dicendi* verbs: it is the only encyclopedia in which (i) *devoir* is associated with none of the *dicendi* verbs and (ii) *falloir* is only specifically associated to *dire*.

Figure 7:

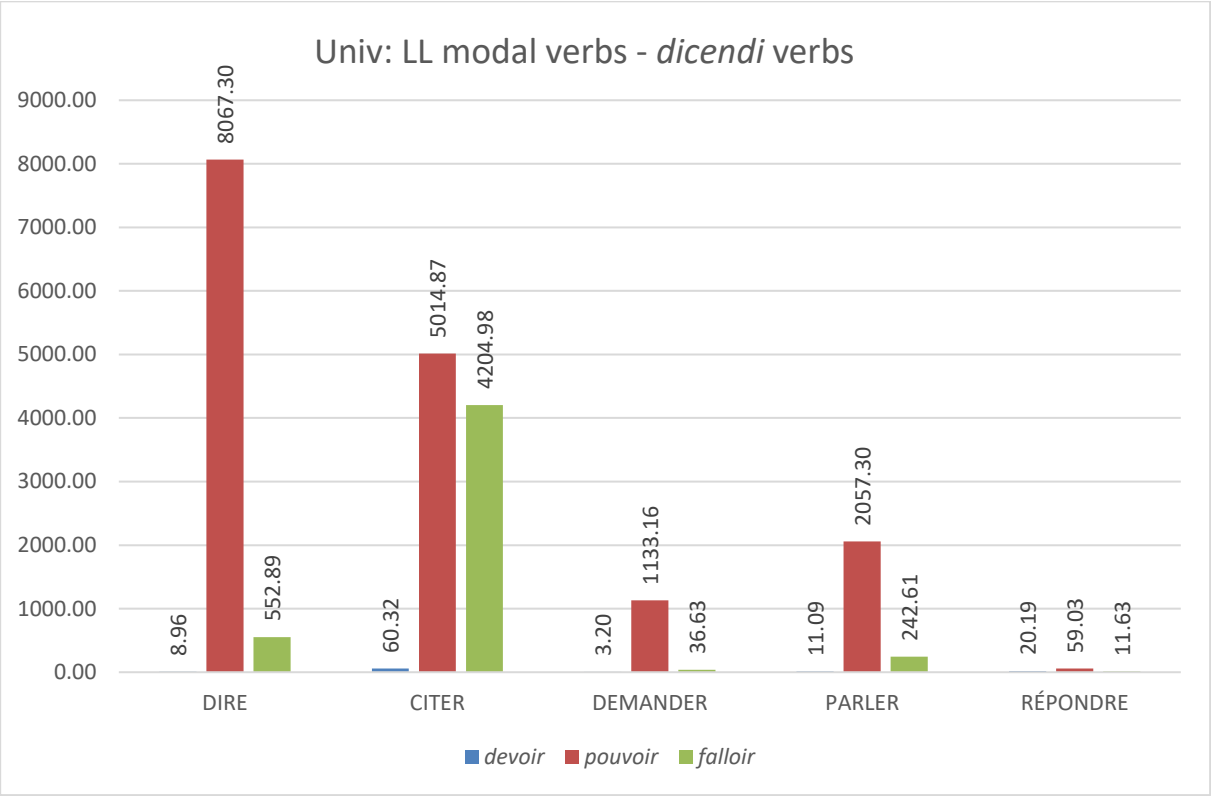
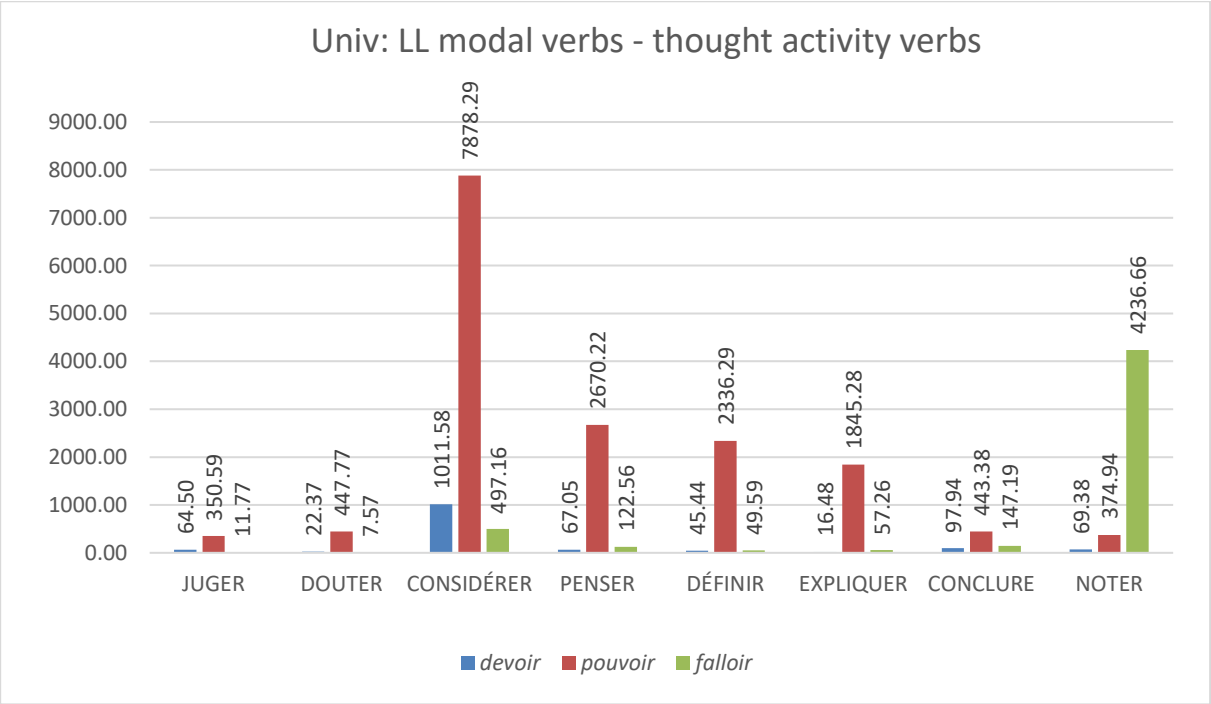


Figure 8:



Concerning *Universalis*, we can notice a particularly strong attraction between *falloir* and *citer* and between *falloir* and *noter* (note that the latter is the only case in which *falloir* is more strongly associated to a lexical item than *pouvoir*), we can hypothesize that such an association is on its way to becoming a pattern in the sense of a pre-constructed sequence: *falloir noter* and *falloir citer*[2]. As for *pouvoir*, this

[2] In spoken corpora, the pattern *faut croire* used as a reaction to a question: - *Il est parti? - Oui, faut croire, sa voiture n'est pas dans le parking* ('- Has he left? - Yes, I guess he must have. His car is not in the parking lot')

encyclopedia shows a strong increase in its association with all *dicendi* verbs except for *répondre*.

Figure 9:

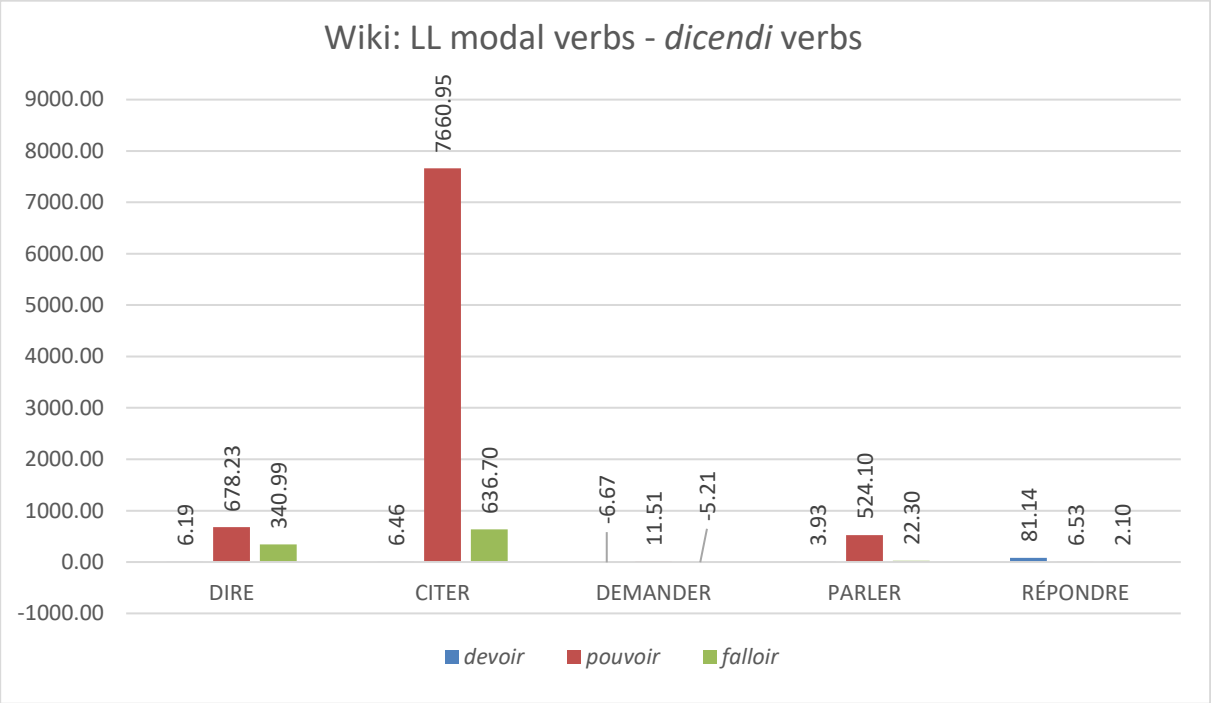
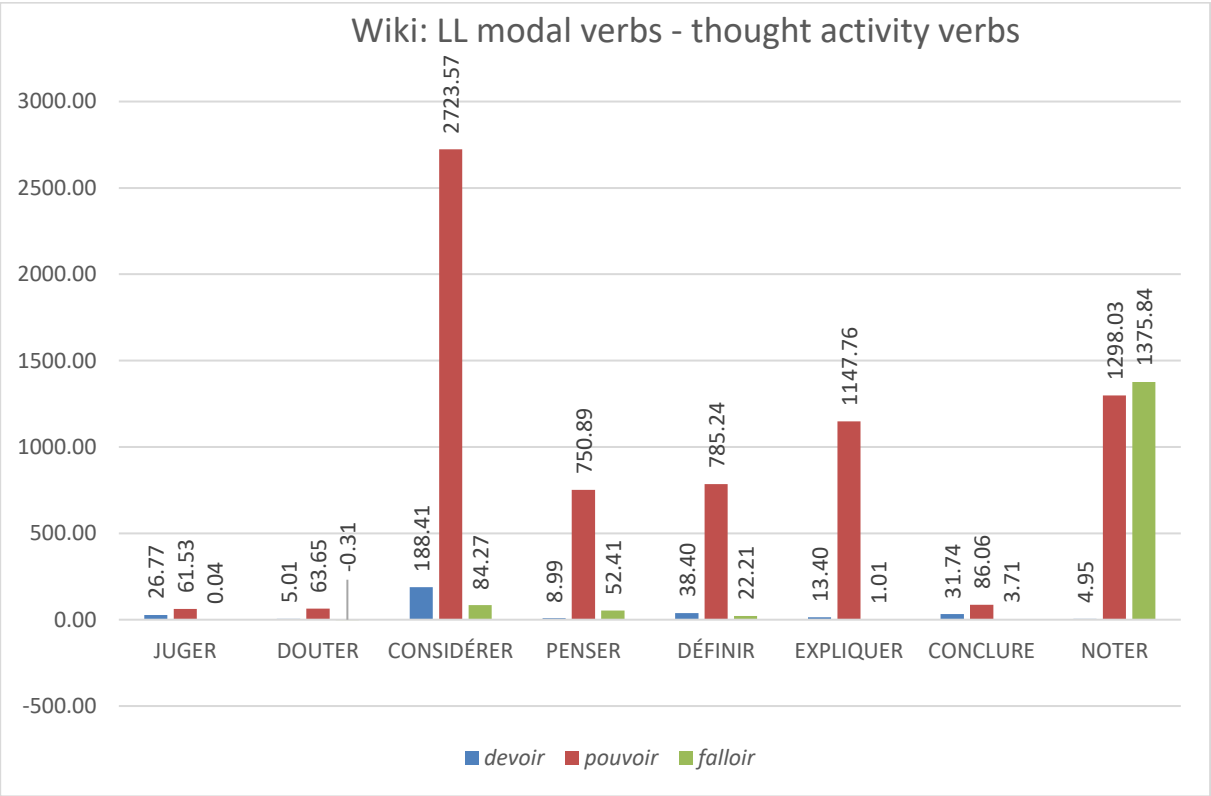


Figure 10:



In *Wikipédia*, the pattern *falloir* + *noter* emerges even more evidently – although the LL value of the association is weaker – since the association of *falloir* with all the other thought activity verbs is declining.

corresponds to a fixed idiom (see Rossari 2012 for a qualitative analysis). This means that *falloir* can combine with thought verbs and forms more or less fixed patterns with some of them.

Overview:

Pouvoir and *falloir* show consolidating patterns: *pouvoir* is already significantly associated to *citer* in *Diderot et D'alembert* but this association strongly increases in both 21st century encyclopedias. The association score between *falloir* and *dire* is significant in *Diderot et D'Alembert*, but it does not change sensibly in the other encyclopedias, whereas the associations *falloir* + *noter* and *falloir* + *citer* strongly emerge in the 21st encyclopedias, revealing a pattern.

3.2 LL measures for morphological modality

We have used the same methodology to select the nouns associated with the *-able* or *-ible* adjectives. We have selected the first 25 nouns (collocates) according to their rank assigned by the LL value in each encyclopedia. We then have divided them into three semantic categories: (i) abstract nouns, (ii) concrete nouns, and (iii) measure nouns. Each graph shows the nouns specific to these suffixes for each category in the three encyclopedias.

The results reveal three tendencies. First, there are disparities between concrete and abstract nouns on the one hand, and measure nouns on the other hand, the latter being preferred in *Universalis* and the former in *Diderot et D'Alembert* and *Wikipédia*. Second, there are important discrepancies between *Universalis* and *Wikipédia* concerning concrete nouns: the same word can show very different LL values for *Universalis* and for *Wikipédia*. Third, as for the measure nouns, which should be stable in theory (there is no reason why they would be differently assessed in terms of the suffixes *-able/-ible*), surprisingly, there is little overlapping between encyclopedias, in particular comparing *Diderot et D'Alembert* with both the 21st century encyclopedias.

Figure 11:

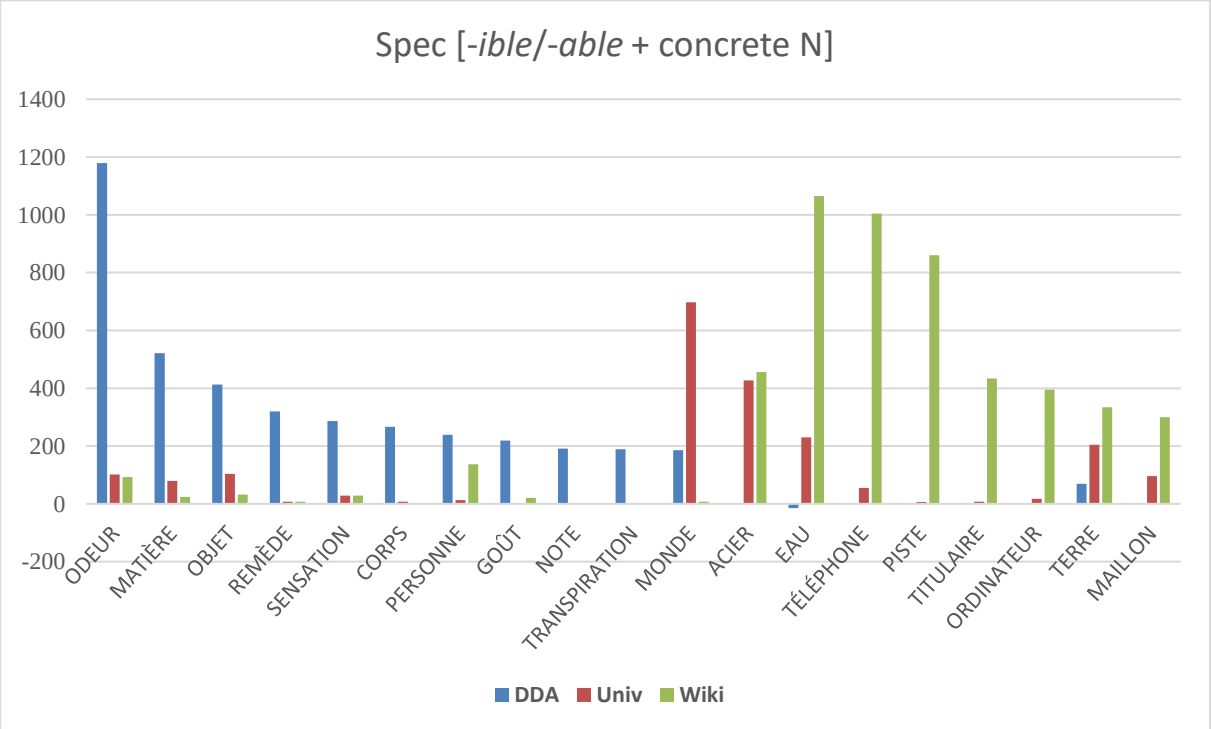
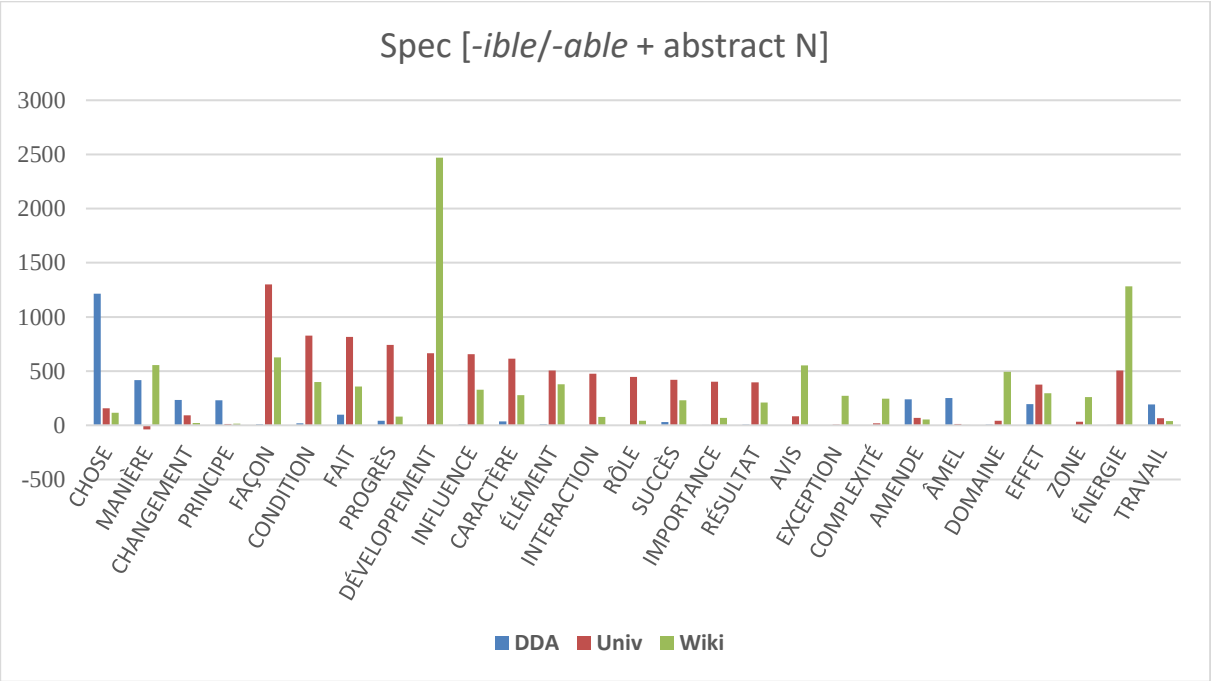


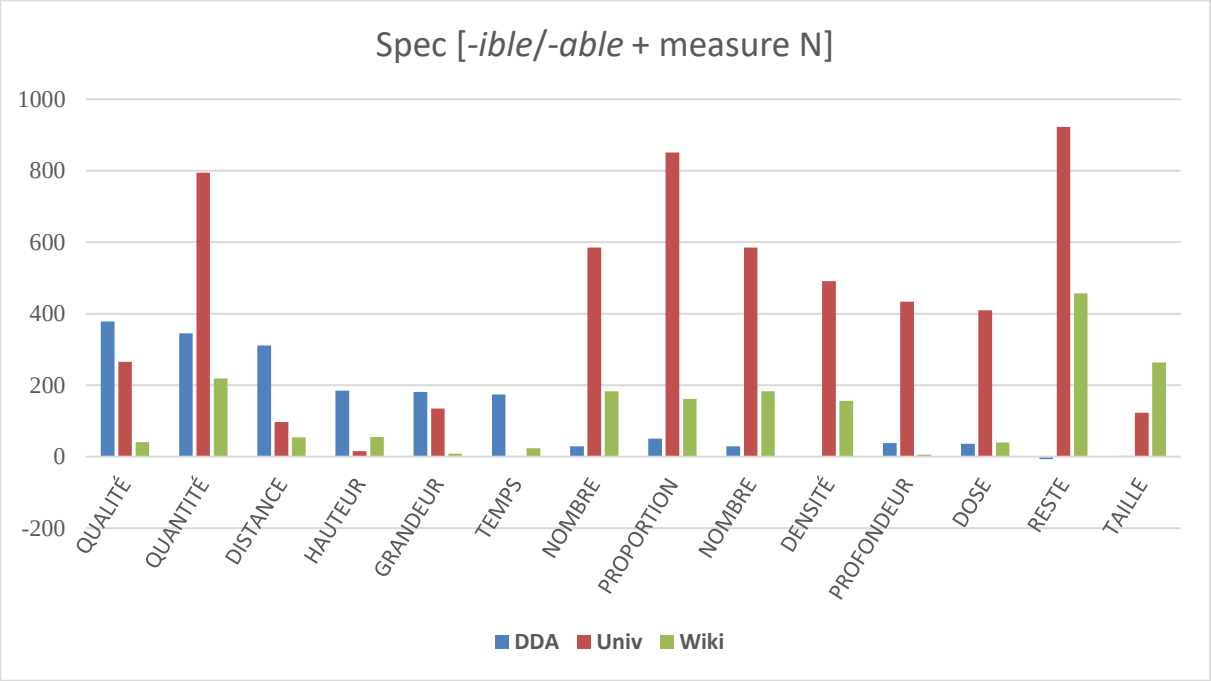
Figure 11 shows the proximity between *Diderot et D'Alembert* and *Wikipédia*. Concrete nouns show strong associations in *Diderot et D'Alembert* and *Wikipédia* (even if, as we can expect, the nouns are different) and weaker associations in *Universalis*. In *Diderot et D'Alembert*, the concrete nouns represent sensations, whereas in *Wikipédia* they refer to day-to-day objects. But there are also differences (that we have no reason to expect) between *Universalis* and *Wikipédia*: the specificities concern much more usual and frequent nouns – such as *ordinateur*, *telephone* – in the latter, which is due to the compound noun including the adjective *portable*.

Figure 12:



Abstract nouns which, theoretically, should be present in the three encyclopedias (*façon*, *condition*, *fait*, *progrès*...) show strong attraction with *-ible* and *-able* suffixes in *Universalis* and, to a lower extent, in *Wikipédia*, but not in *Diderot et D'Alembert*. The strong association with *développement*, which contrasts with the other words in *Wikipédia*, is due to the word pattern 'développement durable', frequently used in this encyclopedia.

Figure 13:



The little similarity between *Diderot et D'Alembert* and *Universalis*/*Wikipédia* concerning measure nouns is due to the presence of scientific terms: *proportion*, *nombre*, *densité*, *profondeur*. They are associated with *-able*/*-ible* in *Universalis* (and to a lower extent in *Wikipédia*), whereas *qualité*, *distance*, *hauteur*, *grandeur*, *temps* are associated with *-able*/*-ible* in *Diderot et D'Alembert*.

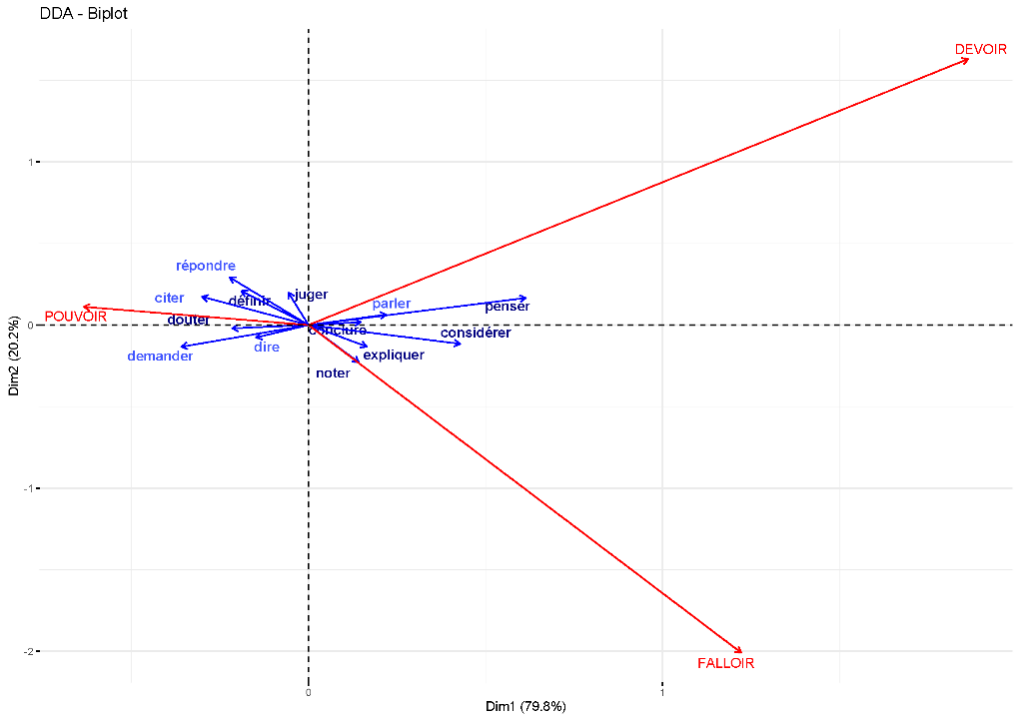
Overview:

The similarities and divergences concerning the use of these categories of nouns indicate clear-cut differences between the encyclopedias, some of which go beyond the diachronic spans, showing the proximity between *Wikipédia* and *Diderot et D'Alembert* for the use of concrete words.

4. Correspondence analysis for lexical modality

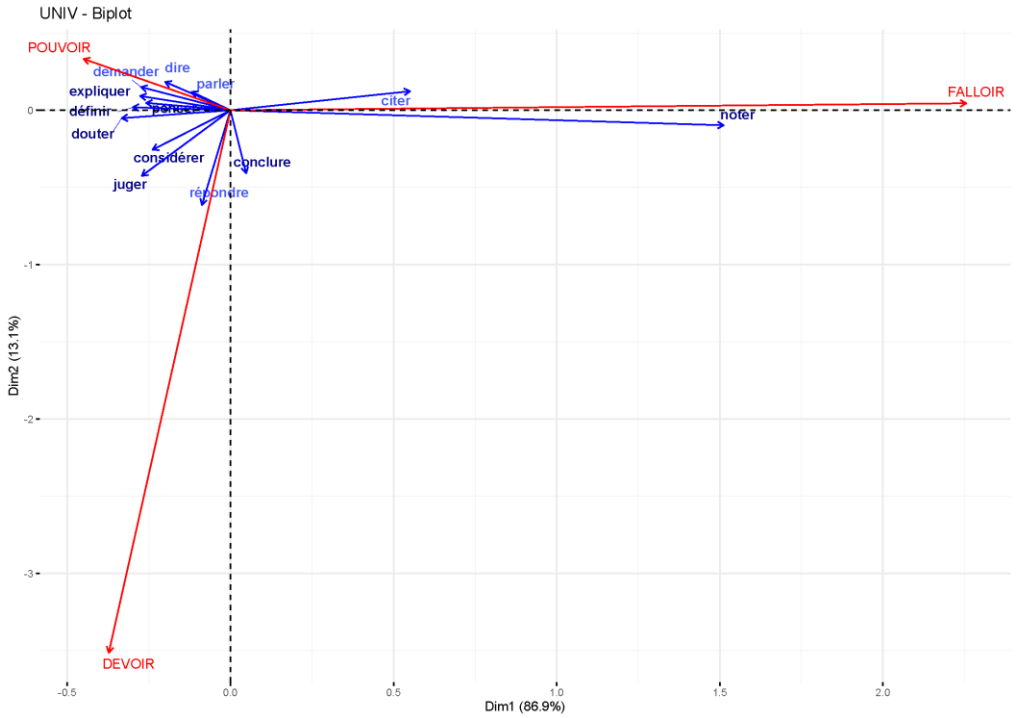
We have applied correspondence analysis (CA) to the verbal lexical forms of necessity and possibility and to both the categories of *dicendi* and of thought verbs. This method is based on a contingency table containing the frequencies of two qualitative variables: in our case, modal verbs (in red) and cognitive verbs (in light blue for *dicendi* verbs and dark blue for thought verbs). The CA results are graphically represented below in the form of asymmetric biplots, which allows us to measure distances between forms representing different variables (e.g. between a modal verb and a cognitive verb), as well as distances between forms representing the same variable (e.g. between two modal verbs or two cognitive verbs). We can thus (i) observe groups of forms and interpret these groupings as indicators of similar behavior; (ii) weigh the results of the LL values against the grouping emerging from the CA, noticing for instance that some LL values showing attraction between two items are less relevant if other forms are taken into account, as it is the case in CA.

Figure 14:



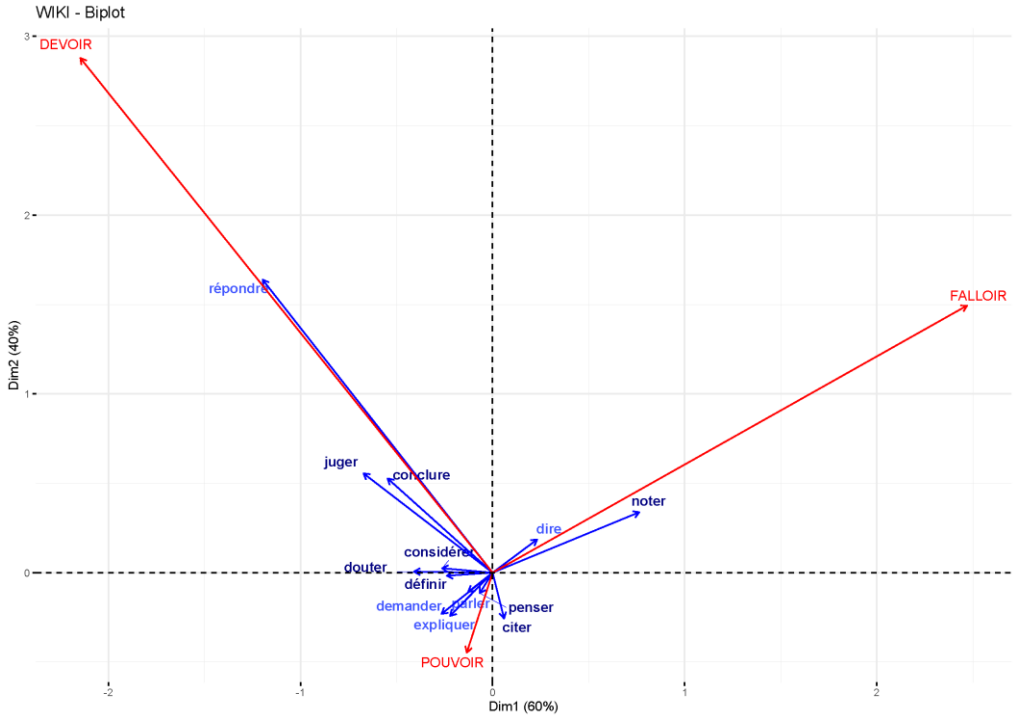
The biplot for *Diderot et D'Alembert* shows (i) clear differences between some thought activity verbs and some *dicendi* verbs: *penser* and *considérer* are close both to *devoir* and to *falloir*, whereas most *dicendi* verbs, namely *citer*, *dire*, *répondre*, and (with the strongest association) *demander* are close to *pouvoir*. However, the distance from the origin of both *devoir* and *falloir* and their associated verbs is greater than that of *pouvoir* and its associated verbs. This can be interpreted as a sign of a more particularized behavior of thought activity verbs compared to *dicendi* verbs; (ii) a proximity between the two necessity verbs: they are closer to each other than they are to the possibility verb.

Figure 15:



The proximity seen above between necessity verbs cannot be observed in the biplot of *Universalis*. Other particularities emerge: (i) *juger* and *répondre* form a group in relation to *devoir*, and *citer* and *noter* in relation to *falloir*; (ii) *pouvoir* remains close to the origin, showing no particular behavior in relation to cognitive verbs.

Figure 16:



The biplot for *Wikipédia* shows similarities with *Universalis*: (i) there is no proximity between necessity verbs; (ii) the group formed by *juger* and *répondre* in relation to *devoir* remains; one important difference should be noted: only *noter* (and not *citer*) is close to *falloir*. In this case, the CA weighs what is observed with the LL values: the score for *falloir* + *citer* is interpreted as highly relevant (cf. 2.1, overview), as it is

largely over the specificity threshold (10.83), but the CA shows that this score does not particularize the behavior of *citer* with respect to the association of the other cognitive verbs with *falloir*.

Overview:

There are one common point and one difference between the three encyclopedias: *pouvoir* is always close to the origin, which means that it shows no particular behavior in relation to cognitive verbs; on the other hand, they diverge in relation to necessity verbs, with *Diderot et D'Alembert* showing proximity between them, which can no longer be observed in the 21st century encyclopedias.

The interpretation of some associations highlighted by LL scores as patterns should be rethought: only *noter* + *falloir* stands out as a pattern.

5. Conclusion

Crossing statistical methods and data was our means to assess different uses of modality across our three encyclopedias. We have applied the same methodology used in Rossari et al. (2018) with the same corpora, which had already led to conclusive results concerning argumentative patterns involving epistemic adverbs and connectives. This time, the statistical approach has newly revealed that the modality of necessity and that of possibility are used differently in a 18th-century encyclopedic corpus compared to two 21st-century encyclopedic corpora. Not only does this methodology indicate differences concerning over- and under-representation of certain modal forms, in particular encyclopedias, but it also reveals their different exploitations depending on their association with lexical items.

References

Blumenthal P. (2017). Evolution de la combinatoire prépositionnelle : le cas de *en*. In Badiou-Monferran, C., Bajric, S. and Monneret, P. editors, *Penser la langue. Sens, texte, histoire*. Paris: Honoré Champion, pp. 161-176.

Charolles M., Diwersy S. and Vigier D. (2017). Evolution des emplois des marqueurs de topiques de discours dans Le Figaro de la fin du XIXe et du début du XXIe siècles. *Langage*, 206: 85-104.

Desagulier G. (2017). *Corpus Linguistics and Statistics with R, Introduction to Quantitative Methods in Linguistics*. Cham: Springer International Publishing.

Evert S. (2008). Corpora and collocations. In Lüdeling, A. and Kytö, M. editors, *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, pp. 1212-1248.

Greenacre M. (2017). *Correspondence Analysis in Practice*. 3rd Ed. Boca Raton, Florida: CRC Press.

Rabatel A. (2003). Les verbes de perception en contexte d'effacement énonciatif : du point de vue représenté aux discours représentés. *Travaux de linguistique*, 46: 49-88.

Rossari C. (2012). Valeur évidentielle et/ou modale de *faut croire*, *on dirait* et *paraît*. *Langue française*, 173 (1): 65-81.

Rossari C., Dolamic L., Hütsch A., Ricci C. and Wandel D. (2018). Discursive Functions of Epistemic Adverbs: What can Correspondence Analysis tell us about Genre and Diachronic Variation? *Actes JADT 2018*, Roma: UniversItalia, pp. 668-675. http://lexicometrica.univ-paris3.fr/jadt/JADT2018/actes-jadt18.pdf

Vigier D. (2017). L'évolution des usages des prépositions *en*, *dans*, *dedans* entre le XVIe et le XXe siècles : approche distributionnelle sur corpus outillé. *Discours* [En ligne], 21, https://journals.openedition.org/discours/9373

Corpora

All the corpora used were supplied by the platform BTLC (Base Textuelle Lexico-statistique de Cologne – http://persan.rom.uni-koeln.de/btlsc), and were constituted within the French-German projects Presto (http://presto.ens-lyon.fr) and Emolex (http://emolex.u-grenoble3.fr). The following corpora have been used:

Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers (ARTFL), version 3, 23 940 181 tokens,

Encyclopædia Universalis (ed. 2005), Presto version, 49 859 864 tokens,

French *Wikipédia* (ed. 11.06.2015, 1 out of 11 articles), Presto version, 50 396 345 tokens.