Règles de normalisation orthographique (niveau 0 original => niveau 1 orthographe normalisée)
MàJ 07/03/2016

01. Singulier/pluriel
01.a si toutes les occurences sont au singulier, ou toutes au pluriel, on ne touche à rien.
01.b si au moins 75% des occurences sont au au pluriel, on met au pluriel les 25% restant (75/25 arrondis, 74,50 = 75). Sinon, on laisse au singulier.
§ si une forme est ambigue (e.g. *agrémens*, peut être singulier), elle n'est pas prise en compte dans le calcul. Si une forme est mal orthographiée, elle n'est pas prise en compte dans le calcul non plus, sauf si c'est l'unique occurrence du substantif.
§ Pour une ligne sans indication de nombre d'occurrences dans la base, on compte comme une seule occurrence.
§ Le calcul se fait sur les catégories lorsqu'elles existent (début d'article): *et habillement, et habillements*, *d'habillement* ou *d'habillements* ne sont pas comptés dans le calcul, ce sont les catégories principales *Habillement* et/ou *Habillements* qui existent et sont prises en compte.
§ Lorsque le genre d'un substantif est ambigü (*bois*), le calcul se fait sur les adjectifs qualificatifs. En leur absence, ou en cas de doute, le substantif est considéré comme singulier.
§ Lorsqu'une forme existe à la fois comme catégorie principale de marchandise et comme substantif dans une énumération et/ou qualificatif, le nombre déterminé sur la base de la forme principale s'applique aussi aux substantifs énumérés et qualificatifs (ici par exemple si *Habillement* est au singulier, *et habillements* sera mis au singulier même si *et habillement* n'existe pas dans l'original); sauf si le qualificatif désigne la matière (*de vache* reste au singulier même si le substantif *vaches* est au pluriel), ou est utilisé comme adjectif d'un autre substantif au singulier (dans *toile mousseline*, *mousseline* n'est pas mis au pluriel même si *mousselines* substantif est au pluriel). En cas de doute sur le statut de matériau, par exemple s'il s'git d'un génitif qui peut être compris comme partitif (*cuisses d'oies*, *cornes de vaches*) c'est-à-dire désignant une partie d'un animal plutôt que la matière, on met au singulier (*cuisses d'oie*, *cornes de vache*).
§ Lorsqu'une forme n'existe que dans des énumérations (*cobalt* dans *azur et cobalt*), ou uniquement comme qualification (*de Galle*, *en gabares*), le calcul se fait sur le nombre d'occurrences dans les catégories principales séparées de la base "normalisation orthographique", indépendamment du nombre d'occurrence dans la base principale à l'intérieur chacune de ces catégories.
[Cette règle pourrait être revue via un traitement sous Stata, mais impossible de faire autrement sous Excel sauf à construire une macro VBA très complexe]
01.c On met au singulier le substantif à l'usage mixte et les adjectifs qui en dépendent directement, c'est-à-dire ceux dont le genre est gouverné par le genre du substantif (évidemment... *arçons à selle* possible lorsque l'on a *à selle* dans certains cas et *à selles* dans d'autres, mais pas *arçons sellé*, sans s). Mais cette règle ne s'applique pas aux qualificatifs indirect (*de modes*, *en feuilles*, par exemple: *papiers en feuilles => papier en feuilles*)
01.d Les mêmes règles s'appliquent aux qualificatifs, etc. en cas de multiplicité des formes SAUF
01.e Si plusieurs substantifs sont enchaînés, on respecte les singuliers / pluriels des catégories, même si cela conduit à une certaine incohérence du nombre dans l'énumération.
01.f Si une forme singulier ou pluriel n'est pas attestée par les dictionnaire, alors qu'elle est dominante dans la base, mais sur un petit nombre d'occurrences et sous des formes majoritairement corrompues, la forme du dictionnaire est insérée
Alibania, alibanic, alibanir, alibanis dans la base ; alibanies dans tous les dictionnaires => alibanies

01.g si un qualificatif d'un substantif apparaît ailleurs comme substantif remplaçant ce substantif, son nombre est celui du substantif principal.
tortues au pluriel, => *tortues cahouannes => cahouannes* lorsque *cahouannes* apparaît seul.

02. Tout est mis en minuscules sauf les noms propres [=> liste des noms propres maintenue à part]; en cas de doute, on met en minuscule.

03. Unification des formes orthographiques différentes: unifier uniquement a) si les occurrences d'une des deux formes à unifier sont inférieures à trois; b) si les deux formes à unifier relèvent de l'un des cas suivants:
N.B.: Si plusieurs formes de la base sont attestées dans les dictionnaires, on garde celle qui se trouve dans la base et qui est la plus fréquente; en cas de doute (moins de 75%/25%), on garde la première dans l'ordre alphabétique.
ajanis, janis => ajanis
03.a La différence vient d'une différence singulier/pluriel attestée dans les dictionnaires *ails = aulx*
03.b Le nombre de syllabe est identique (e muet compté comme une syllabe) et les phonèmes sont identiques dans l'ordre sur les deux types d'occurences
acier = assié
03.c Les phonèmes peuvent être considérées comme identiques si les deux conditions 03.c.1 ET 03.c.2 sont remplies:
03.c.1 consonnes de même type d'une des deux façons suivantes
- soit phonétiquement (occlusives vs fricatives.)
*a**l**quifoux = a**r**quifoux*
- soit proches par la forme de la lettre
*alpargal**l**e = alpargat**t**e*
03.c.2 voyelle identique (hors diphtongue)
***a**r**a**cq = **a**r**a**k, **a**lp**a**rg**a**lle = **a**lp**a**rg**a**tte*
03.d Les phonèmes finaux sont considérés comme identiques si terminaison latine / terminaison moderne (e / us); on garde la forme la plus fréquente à 75%; si doute, garder la forme latine:
antimoine = antimonium => antimoine si antimoine est plus fréquent, *antimonium* autrement.
03.e La seule différence entre les deux occurrences est le placement des espaces:
alqui foux = alquifoux
03.f La seule différence entre les deux occurrences est l'absence d'une lettre ou d'une syllabe
03.f.1 Une seule lettre est différente, elle appartient à un groupe de lettres comparables se resssemblant par les jambages et la taille, comme suit (a/c/e/i/o/r/s/u/v/x; b/d/h/l/t; f/h/l/S/; a/i/n/r/s/u/v), et le nombre d'occurrences est inférieur à trois
bas de Saint-Meant = bas de Saint-Mexant
cranettes des Indes = canettes des Indes
03.f.2 Une seule syllabe manque, mais les phonèmes sont parfaitement identiques (et pas seulement considérées comme identiques), le mot est de quatre syllabes ou plus, le nombre d'occurrences est inférieur à trois, le mot n'est attesté nulle part, et la même syllabe apparaît comme manquante en lecture automatisée > présomption de mauvaise lecture ou de mauvaise transcription (utiliser Google Books)
artoloche = aristoloche
03.g A l'occasion, essayer de rechercher dans le pdf d'origine une erreur de transcription (*ad lib/wishful thinking*)

04. Unifier ?, ??, ??? et autres expressions de doute en "???"

? brun => ??? brun

05 Moderniser l'orthographe
05.a Utiliser une orthographe moderne chaque fois que cela est possible, y compris en cas de changement de genre
bierre d'angleterre => bière d'Angleterre
Agats fins => agate fine
05.b En cas de cohabitation de formes modernes et anciennes, la normalisation se fait a) en fonction du nombre d'occurrence (75/25 comme singulier/pluriel); b) uniquement sur la forme moderne sauf si c'est du latin
ail + ail + ails + aux + aux + aux + aux => ail

06 En cas de produit inconnu,
06.a Faire d'abord une recherche rapide dans les principaux dictionnaires et sources primaires
N.B.: En cas de résultat, noter référence ET définition (réintégrer le dictionnaire de Loïc comme colonne dle cas échéant)
06.b A l'occasion, essayer de rechercher dans le pdf d'origine une erreur de transcription (*ad lib/wishful thinking*)

07 supprimer les séparateurs (virgules, points-virgules, "/")
Ajustements ; de femme => ajustement de femme

08 développer les abréviations si possible quand il n'y a aucun doute (abréviation développée par ailleurs pour la même marchandise)
joncs à fre des chaises => joncs à faire des chaises

09 unifier les références à des noms de lieu sur la forme moderne quand il n'y a aucun doute (référence développée par ailleurs pour la même marchandise)
bois Brésil, bois de Brésil => bois du Brésil
Quand il y a doute (mouchoir de toile dite "Cholet", ou mouchoir provenant de la ville de Cholet; étoffe dite "Bengale", ou en provenance du Bengale), laisser la forme d'origine.