Méthode d'indexation par les multi-termes

Bernard DOUSSET
IRIT/ SIG, Université Paul Sabatier,
118 , route de Narbonne, 31062 Toulouse cedex 04
dousset@irit.fr

1. Problématique

Lorsque nous voulons analyser sémantiquement un corpus hétérogène dont l'indexation repose sur des thésaurus spécifiques à chaque source, il est illusoire de vouloir fusionner les différents champs d'indexation. En effet, non seulement les mots-clés ne sont pas à jour, mais ils typent les documents en fonction de leurs sources (vocabulaires d'indexation quasiment disjoints). La solution que nous préconisons est de réaliser une indexation automatique et homogène des champs en texte libre (titre, résumé, texte intégral) s'appuyant sur un thésaurus à jour du domaine étudié et qui est généré par l'analyse syntaxique, sémantique et statistique de la totalité des champs traitant du sujet sur l'ensemble des sources. Ainsi, un mot-clé (venant d'une base ou d'un auteur) peut servir d'index à tous les documents du corpus.

2. Notion de multi-termes

La terminologie rencontrée dans les champs sémantiques peut se décomposer en trois entités :

- Les mots simples ou uni-termes : le dictionnaire de la langue au sens propre du terme.
- Les radicaux auxquels il est possible de ramener certains uni-termes par des algorithmes de radicalisation (stimming, algorithme de Porter, …).
- Les mots composés, syntagmes ou expressions : suite ordonnée d'uni-termes comme : « analyse de données », « état de l'art ». Ils peuvent éventuellement être reliés par des tirets et/ou suivis de leur acronyme comme « bovine-spongiform-encephalopathy (BSE) » ou « Creutsfeld Jacob desease CJD ».

Ces derniers sont bien entendu beaucoup plus précis et leur valeur sémantique leur permet de faire office de mots-clés. Aussi, doit-on les rechercher systématiquement dans les textes et, si possible, en générer un dictionnaire qui va servir de base à l'indexation automatique et à l'analyse sémantique.
Habituellement nous procédons comme suit :

- Générer les dictionnaires de tous les champs sémantiques (thesaurus, mots-clés, index, classifications, termes d'indexation des auteurs, titres, résumés, texte intégral, …).
- Fusionner ces dictionnaires et dédoublonner.
- Ne garder que les mots composés sans leurs acronymes,
- Générer un dictionnaire de multi-termes de la spécialité (suite d'uni-termes séparés par des espaces).
- Eventuellement générer un dictionnaire de synonymes notamment pour prendre en compte les acronymes et éventuellement les variations morphologiques (inversion, terminaisons, pluriels, multilinguisme, …).

Cette première phase permet d'extraire, d'un volumineux corpus, tout l'aspect conscient de l'information dite explicite qui comprend :

- Le conscient collectif : terminologie sur laquelle tout le monde est d'accord et essentiellement représentée par la notion de mots-clés.
- Le conscient individuel : mots composés signalés par certains auteurs et détectables par la présence de tirets et/ou d'acronymes.

Mais il s'agit d'un vocabulaire convenu qui n'a pas une grande utilité pour détecter l'innovation et les sujets tout juste émergents. Il sert toutefois à parfaitement cibler les grands axes du domaine et éventuellement leurs interconnexions.

3. Détection d'une nouvelle terminologie

Dans une seconde phase, nous allons rechercher l'ensemble des multi-termes qui ne sont signalés par personne, car totalement nouveaux dans le domaine, et qui représentent, à nos yeux, le front de recherche (ou d'innovation) encore inconscient, mais qui se trouve à l'état de traces dans les textes que nous analysons.

Pour cela, nous recherchons de façon statistique quelles sont les expressions qui reviennent suffisamment souvent (au moins deux fois) et qui sont absentes du dictionnaire « conscient » précédent. Cette détection nous permet d'accéder à une information qui n'est ni pointée par les mots-clés traditionnels, ni proposée par les auteurs, car trop récente ou non encore officiellement reconnue.

Elle présente elle aussi deux niveaux bien distincts :

- L'inconscient collectif : une même expression se retrouve chez plusieurs auteurs, mais personne ne la signale comme mot-clé éventuel (ces auteurs ont donc dû lire ou écouter un message qui les a séduits, ils sont d'accord mais ne le savent pas encore. Il y a donc un consensus inconscient).
- L'inconscient individuel : un même auteur utilise plusieurs fois dans ses écrits une nouvelle expression, il s'agit d'un « segment répété » qui représente l'idée importante de son discours.

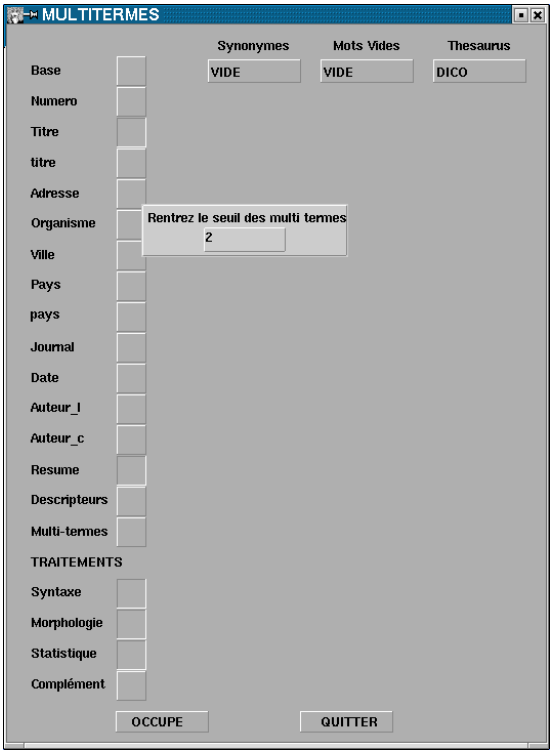

Figure 1. : Génération des multi-termes

Bien entendu, ces deux niveaux d'information sont beaucoup plus subtils que les précédents, ils nous permettent d'accéder à la nouveauté ou à tout ce qui touche au marginal. Leurs croisements avec tous les autres éléments d'information vont nous permettre de savoir quels sont les acteurs concernés, dans quels axes apparaissent ces nouveaux concepts et éventuellement s'ils s'inscrivent dans des stratégies visibles. Les propositions faites par cette méthode automatique de détection doivent être validées en deux temps :

- Tout d'abord, par les documentalistes, afin d'éliminer certaines expressions usuelles de la langue et des éléments trivialement inutiles.
- Ensuite, par les experts du domaine, afin de ne garder parmi les multi-termes détectés (expressions, mots composés, molécules, sigles, …) que ceux ayant une réelle valeur sémantique dans le cadre de la problématique étudiée.

Mais bien souvent, nous détectons également des évidences qui ont échappé à tout le monde, notamment aux indexations traditionnelles, et qui ont une importance stratégique indéniable.

4. Création d'un champ d'indexation complémentaire

Nous allons générer, dans chaque document, un nouveau champ d'indexation reprenant la présence de la terminologie simple ou composée au niveau des différents champs sémantiques. Cette génération s'appuie sur le dictionnaire validé des multi-termes, sur un dictionnaire de mots vides à éliminer ainsi que sur un dictionnaire de synonymes d'uni-termes. Elle produit un nouveau champ d'indexation (cette fois-ci homogène) qui reprend la liste des éléments détectés. Il est possible d'y adjoindre certaines options :

- Détection statistique pour générer le dictionnaire « inconscient ».

- Traitement syntaxique pour tenir compte de la ponctuation qu'un multi-terme ne peut franchir.
- Traitement morphologique (radicalisation pour éviter d'utiliser de trop gros dictionnaires de correspondance).

Mais outre les quatre niveaux d'information cités plus haut, le principal intérêt de cette indexation est d'être homogène pour l'ensemble des sources utilisées. Il est donc possible, au sein d'une analyse multi-bases, d'utiliser les qualités sémantiques offertes par ces nouveaux mots-clés. De plus, les sources collaborent et donc améliorent mutuellement l'indexation commune, car une expression détectée dans l'une d'entre elles permet d'indexer correctement toutes les autres, de même pour un multi-terme découvert grâce à sa présence sur au moins deux supports différents.

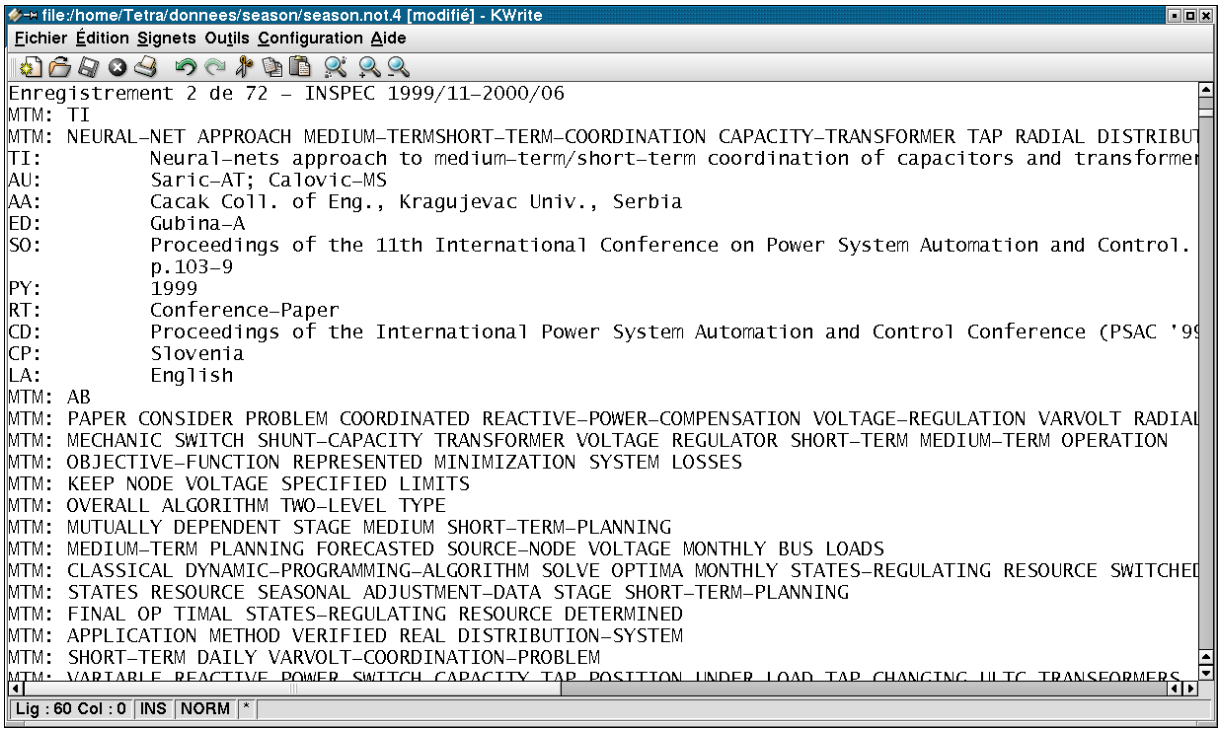

Figure 2. : Ajout d'un nouveau champ d'indexation MTM :

5. Choix de la terminologie optimale

Le principal problème posé par le champ d'indexation complémentaire des multi- termes est la très grande taille du dictionnaire qui permet de le générer. Souvent plusieurs dizaines de milliers d'entrées. Il n'est donc pas toujours possible de charger en mémoire l'ensemble des croisements sémantiques qu'il propose. Nous avons donc pensé à simplifier ce dictionnaire en ne conservant que les termes les plus représentatifs, c'est à dire ceux qui typent le mieux les classes sémantiques du domaine et qui créent le moins possible de connexions non significatives. Pour cela nous éliminons :

- les mots vides de la langue,
- les termes qui ne sont présents qu'une fois dans le corpus (apax),
- ceux qui sont distribués uniformément sur l'ensemble des documents (termes de l'équation de recherche, mots usuels ou trop généraux, …),

Mais ce nettoyage n'est pas toujours suffisant pour ramener le nombre de multi-termes à un volume exploitable en mémoire. Nous avons alors recours à une technique mise au point par un de mes étudiants en DEA [**KANOb98**] et qui consiste à ne conserver que les termes qui ont une forte densité dans certains documents. Pour cela nous calculons le rapport entre densité locale (dans chaque document) et densité globale (sur l'ensemble du corpus) et nous qualifions prioritairement les éléments dont le rapport est important dans au moins deux documents. En abaissant progressivement le seuil de qualification nous pouvons ainsi générer un dictionnaire de la taille désirée. Cette procédure permet de détecter le cœur des classes sémantiques par leur terminologie la plus typique. Les liens

éventuels entre classes sont alors dus à des termes précis et non plus provoqués par des expressions courantes sans grande signification.
Dans la figure suivante nous illustrons le principe de qualification des termes :

- Nous calculons la densité relative d'un terme dans chaque document,
- Le seuil est initialement fixé très haut,
- On abaisse progressivement ce seuil,
- Dès qu'un terme dépasse ce seuil sur au moins deux documents, il est qualifié,
- On arrête le processus de qualification dès qu'un nombre fixé de termes qualifiés est atteint.

Quelque soit le volume imposé aux dictionnaires et aux matrices, on est sûr d'avoir choisi les termes qui génèrent les classes sémantiques les plus précises. Une fois trouvé le noyau de chaque classe, il est toujours possible de requalifier certains des termes écartés afin, notamment, de retrouver le contexte et d'enrichir le réseau sémantique. Il suffit pour cela de réaliser un simple croisement entre la classe formée de termes qualifiés et l'ensemble des autres multi-termes.

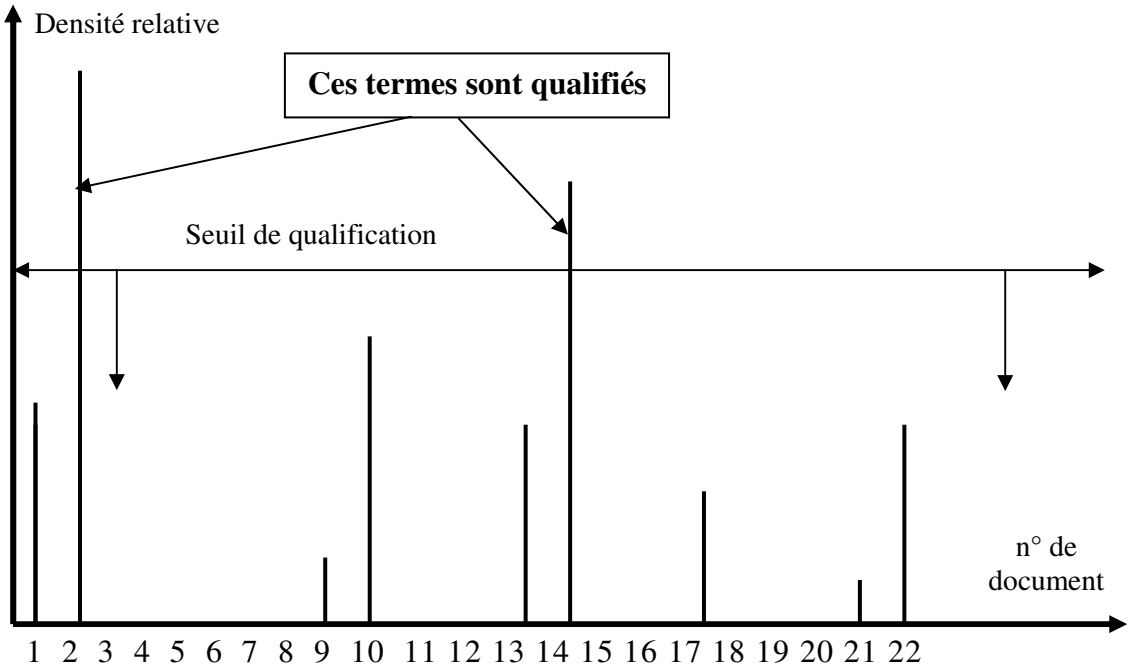

Figure 3. : Qualification d'un multi-terme dans le dictionnaire à conserver

6. Utilisation du champ multi-termes

Un nouveau champ est alors ajouté au corpus, il est composé de la liste des multi-termes retenus et présents dans chaque document indexé. Les multi-termes sont reliés par des tirets et séparés par un espace. Il suffit de décrire ce nouveau champ de la base dans les méta-données pour qu'il soit reconnu par les fonctions d'extraction et de croisement d'information. Il sera alors utilisé comme un champ natif. Son utilité est multiple :

- Il représente un champ d'indexation à jour :
  - conscient collectif (CC)
  - conscient individuel (CI)
  - inconscient collectif (IC)
  - inconscient individuel (II)
- Il est homogène dans un environnement multi-base
- Il peut contenir des termes très spécifique :
  - formules chimiques
  - expressions
  - sigles complexes

Il se substitue donc aux champs d'indexation hétérogènes.

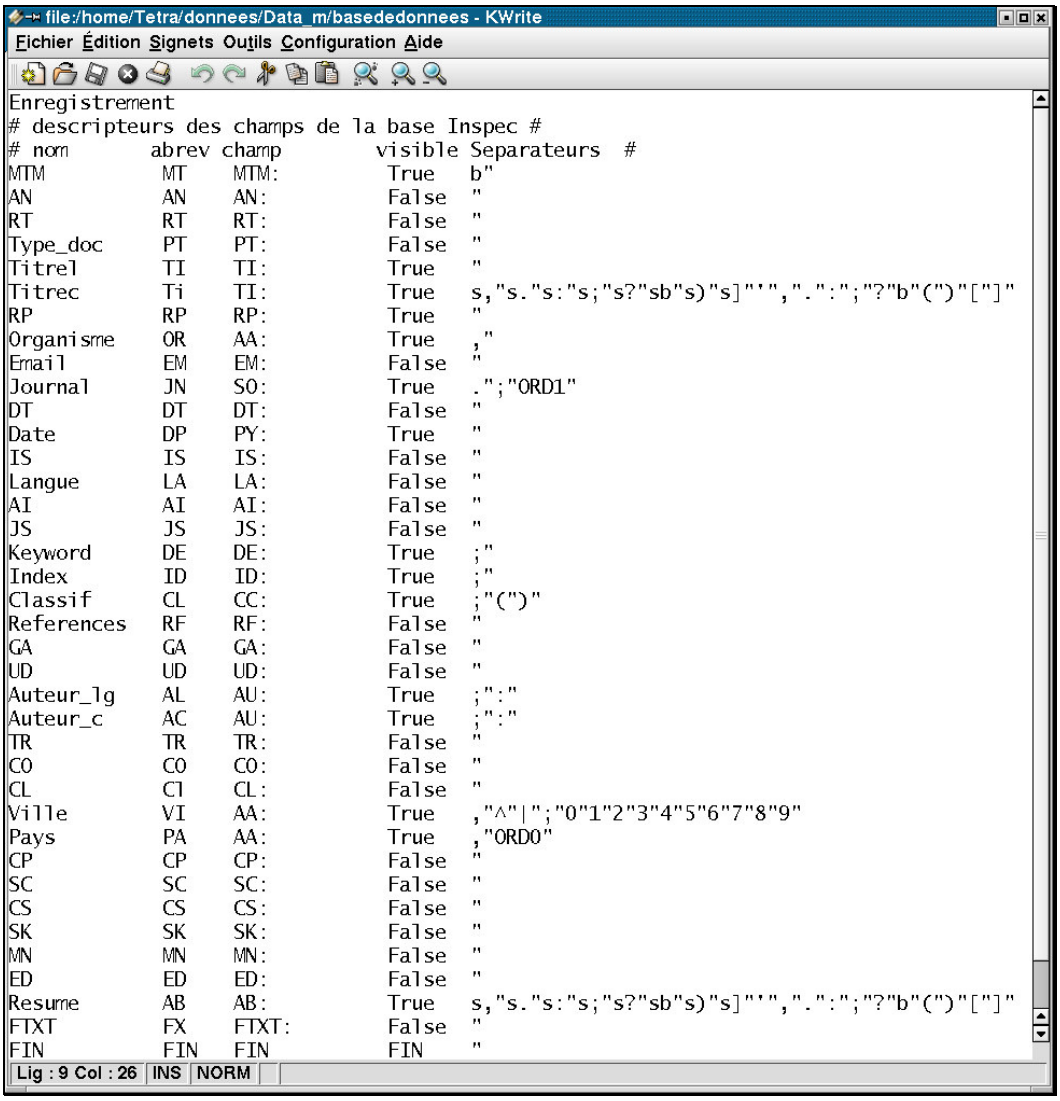

Figure 4. : Prise en compte du champ multi-termes dans les méta-données.

7. Exemple de dictionnaires de multi-termes

Dictionnaire initial (conscients collectif et individuel)

Ce dictionnaire est issu de la compilation des tous les champs d'indexation et de ceux en texte libre qui indiquent des mots composés.

ACID COMMON INORGANIC ANION
ACOUSTIC EMISSION AE
ADJACENT HIGH MELTING POINT
ADSORBED CARBON BLACK
ADSORBED LAYER PVP
ADVANCED ANALYSIS METHOD
AGREEMENT EXPERIMENTAL DATA
ALL ALUMINUM CYLINDER
ALLOY PARTICLE MELT
ALLOY PARTICLE PAD
ALUMINIUM ALLOY FOAM
ALUMINIUM ALLOY SHEET
ALUMINUM ALLOY SHEET
AMINO ACID COMMON INORGANIC ANION
AMPHIPHILIC BLOCK COPOLYMER
ANALYSE CIRCUMFERENTIAL THROUGH WALL

Dictionnaire complémentaire (inconscients collectif et individuel)

Ce dictionnaire est généré par étude statistique sur la redondance dans le corpus d'expressions qui ne sont pas présentes dans le dictionnaire précédent.

hydrogen-trapped
plastic-non-symmetric
load-displacement
molecular-building-block
organic-inorganic
polyurethane-foam
poly-paraphenylene
iron-fibre
oxygen-facilitated
molecular-crystal
ethyl-cellulose
aluminium-alloy-foam
accurate-determination
knitted-fabric-composite
autofrettaged-pressure-vessel
solvent-mediator
temperature-reduction
shape-recovery
sol-gel-process
porphyrinosilica-template
iron-porphyrinosilica
sandwich-shell
foamed-metal
perfect-shell
sheet-yielding
optimally-designed
elastic-deformation
functionally-graded
integrated-process
sixth-conference-hole-burning-related-
platinum-tetra-pentafluorophenyl-porphine-pttfpp
temperature-dependence-atm
pressure-temperature
aqueous-solution-poly

8. Extraction de clusters sémantiques

Voici quelques clusters sémantiques obtenus à partir du croisement des multi-termes les plus typés et tri de la matrice de cooccurrence obtenue par blocs diagonaux :

CLUSTER 1
composite-beam
thermal-expansion
embedded-shape-memory-alloy
thermal-buckling
shape-recovery
temperature-reduction
laminated-composite

CLUSTER 2
solvent-polymeric
liquid-chromatographic
metalloporphyrin-based
porphyrin-based
anion-selectivity
potentiometric-selectivity
lipophilic-anionic
membrane-electrode
membrane-electrode-based
anion-selective
ion-selective-electrode
paper-presented
mnppix-psp4
quaternary-ammonium
excellent-selectivity
cationic-site

CLUSTER 3
manganese-tetraphenylporphyrin
serum-sample
show-high-selectivity-histamine-amino-acid-
common-
carrier-poly-vinyl-chloride-membrane-potentiometri
histamine-synthetic
near-nernstian-response-concentration-range-detect
surface-graphite-electrode
applied-determination

CLUSTER 4
chain-transfer-agent
controlled-molecular-weight
immortal-polymerization
lewis-acid
narrow-molecular-weight-distribution
turnover-number
polymerization-methyl-methacrylate
aluminum-porphyrin-initiator
living-anionic-polymerization
growing-specy
aluminum-porphyrin
acid-assisted
anionic-polymerization
proceeded-rapidly
living-polymerization

ring-opening-polymerization
molecular-weight-distribution
ester-group
micellar-aggregate
two-stage

CLUSTER 5
catalytic-oxidation
development-biomimetic-oxidation-catalysis
structural-feature
thiolate-ligand
n-oxide
drug-metabolism-study
polypeptide-bound
ruthenium-porphyrin
iron-prophyrin
polymer-complex
recent-study
high-yield
extremely-high
one-step
porphyrin-complex
high-selectivity

9. Conclusion

Cette méthode d'indexation s'impose dans le cas d'analyses multi-bases afin d'homogénéiser les termes du vocabulaire d'indexation et ainsi d'harmoniser les contributions de chaque document. En effet, les formats peuvent être très différents ainsi que la qualité et la disparité de l'indexation initiale quand elle existe. Dans le cas d'une seule base, son intérêt reste grand car les indexations proposées ne sont pas à jour et il est alors très difficile de détecter des signaux faibles. Enfin, certains termes techniques ne sont en général pas indexés comme, notamment, les formules de chimie ou des expressions très longues sans compter la possibilité de détecter le copier-collé opération qui peut être pleine d'enseignements.