Lens effects in autonomous terminology learning with conceptual vectors

Mathieu Lafourcade
LIRMM
Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
MONTPELLIER - FRANCE.
lafourca@lirmm.fr
http://www.lirmm.fr/~lafourca

Abstract

In the framework of research in meaning representations for NLP, we focus our attention on thematic aspects and conceptual vectors. The learning strategy of conceptual vectors relies on the morphosyntactic analysis of human usage dictionary definitions linked to vector propagation. The vector of concepts are built from already defined ontologies. We discuss the various effects of translating vectors from a general and coarse grained semantic space to one that has been extended to some specific domain.

1 Introduction

Research in meaning representation in NLP is an important problem still addressed through several approaches. The NLP team at LIRMM currently works on thematic and lexical disambiguation text analysis (Laf01), (LP01). Therefore, we built a system, with automated learning capabilities, based on conceptual vectors for meaning representation. Vectors are supposed to encode *ideas* associated to words or expressions. The conceptual vector learning system automatically defines or revises its vectors according to the following procedure. It takes, as an input, definitions in natural language contained in electronic dictionaries for human usage (a similar approach can be found in (BC01)). These definitions are then fed to a morpho-syntactic parser that provides tagging and analysis trees. Trees are used as a guide by a procedure that computes vectors according to tree geometry and syntactic functions. Nevertheless, a kernel of manually indexed terms is necessary for bootstrapping the analysis and (sometimes) manual definitions are needed when already existing ones prove being too difficult or ambiguous. Furthermore, to assist the vector building procedure, we need to refine the obtained vectors with transversal relationships such as synonymy, antonymy, hypernymy and so forth. The concepts that constitute the basis of the representation come from an already existing thesaurus. The components of vectors are directly related to the concepts of the ontological classification provided by the thesaurus. Aside a general thesaurus (from (Lar92)) that covers the entire semantic space, we can use much more detailed ontologies of speciality. In our experiments, we have taken a hierarchy of concepts from (OEC91) where around 2000 leaf concepts about economy are defined. In this paper, we discuss the various effects on vector representation of switching from the general coarse grain space to a domain specific space. We have applied our models to French and the analysis of press release in the field of economy and stock exchange.

2 Conceptual Vectors

We represent thematic aspects of textual segments (documents, paragraph, syntagms, etc.) by conceptual vectors. Vectors have been used in information retrieval for long (SM83) and for meaning representation by the LSI model (DDL+90) from latent semantic analysis (LSA) studies in psycholinguistics. In computational linguistics, (Cha90) proposes a formalism for the projection of the linguistic notion of semantic field in a vectorial space, from which our model is inspired. From a set of elementary notions, concepts, it is possible to build vectors (conceptual vectors) and to associate them to lexical items [1]. The hypothesis that considers a set of concepts as a generator to language has

[1]Lexical items are words or expressions which constitute lexical entries. For instance, ‹*car*› or ‹*white ant*› are lexical items. In the following we will (some what) use sometimes *word* or *term* to speak about a *lexical item*.

been long described in (Rod52) (*thesaurus hypothesis*). Polysemous words combine the different vectors corresponding to the different meanings. This vector approach is based on well known mathematical properties, it is thus possible to undertake well founded formal manipulations attached to reasonable linguistic interpretations. Concepts are defined from a thesaurus (in our prototype applied to French, we have chosen (Lar92) where 873 concepts are identified to compare with the thousand defined in (Rod52)). To be consistent with the thesaurus hypothesis, we consider that this set constitutes a generator space for the words and their meanings. This space is probably not free (no proper vectorial base) and as such, any word would project its meaning on this space.

2.1 Thematic Projection Principle

Let be $\mathcal{C}$ a finite set of n concepts, a conceptual vector V is a linear combination of elements c_i of $\mathcal{C}$. For a meaning A, a vector $V(A)$ is the description (in extension) of activations of all concepts of $\mathcal{C}$. For example, the different meanings of ‹*quotation*› could be projected on the following concepts (the CONCEPT[intensity] are ordered by decreasing values):

V(‹*quotation*›) =
STOCK EXCHANGE[0.7], *LANGUAGE*[0.6], *CLASSIFICATION*[0.52], *SYSTEM*[0.33], *GROUPING*[0.32], *ORGANIZATION*[0.30], *RANK*[0.330], *ABSTRACT*[0.25], ...

In practice, the largest $\mathcal{C}$ is, the finer the meaning descriptions are. In return, the computer manipulation is less easy. It is clear, that for dense vectors[2] the enumeration of the activated concepts is long and difficult to evaluate. We would generally prefer to select the thematically closest terms, i.e., the *neighborhood*. For instance, the closest terms ordered by increasing distance of ‹*quotation*› are:

$\mathcal{V}$(‹*quotation*›) = ‹*management*›, ‹*stock*›, ‹*cash*›, ‹*coupon*›, ‹*investment*›, ‹*admission*›, ‹*index*›, ‹*abstract*›, ‹*stock-option*›, ‹*dilution*›, ...

2.2 Angular Distance

Let us define $Sim(A, B)$ as one of the *similarity* measures between two vectors A et B, often used in information retrieval (Mor99). We can express this function as:

$$Sim(A, B) = \cos(\widehat{A, B}) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

with "$\cdot$" as the scalar product. We suppose here that vector components are positive or null. Then, we define an *angular distance* D_A between two vectors A and B as:

$$D_A(A, B) = \arccos(Sim(A, B))$$

Intuitively, this function constitutes an evaluation of the *thematic proximity* and is the measure of the angle between the two vectors. We would generally consider that, for a distance $D_A(A, B) \leq \frac{\pi}{4}$, (i.e. less than 45 degrees) A and B are thematically close and share many concepts. For $D_A(A, B) \geq \frac{\pi}{4}$, the thematic proximity between A and B would be considered as loose. Around $\frac{\pi}{2}$, they have no relation. D_A is a real distance function. It verifies the properties of reflexivity, symmetry and triangular inequality. We can have, for example, the following angles (values are in degrees):

D_A(‹*profit*›, ‹*profit*›)=0°
D_A(‹*profit*›, ‹*benefit*›)=10°
D_A(‹*profit*›, ‹*finance*›)=19°
D_A(‹*profit*›, ‹*market*›)=28°
D_A(‹*profit*›, ‹*product*›)=32°
D_A(‹*profit*›, ‹*goods*›)=31°
D_A(‹*profit*›, ‹*sadness*›)=65°
D_A(‹*profit*›, ‹*joy*›)=39°

The first value has a straightforward interpretation, as ‹*profit*› cannot be closer to anything else than itself. The second and the third are not very surprising since a ‹*benefit*› is quite synonymous of ‹*profit*›, in the ‹*finance*› field. The words ‹*market*›, ‹*product*› and ‹*goods*› are less related which explains a larger angle between them. The idea behind ‹*sadness*› is not much related to ‹*profit*›, contrary to its antonym ‹*joy*› which is thematically closer (either because of metaphorical meanings of ‹*profit*› or other semantic relations induced by the definitions). The thematic is by no way an ontological distance but a measure of how strongly meanings may relate to each others.

The graphical representations of the vectors of ‹*exchange*› and ‹*profit*› shows that these terms

[2] Dense vectors are those which have very few null coordinates. In practice, by construction, all vectors are dense.

are indeed quite polysemous. Two other terms (‹*cession*› and ‹*benefit*›) seems to be more focused on specific concepts. These vectors are the average of all possible meanings of their respective word in the general Thesaurus. It is possible to measure the level of *fuzziness* of a given vector as a clue of the number of semantic fields the word meaning is related to.

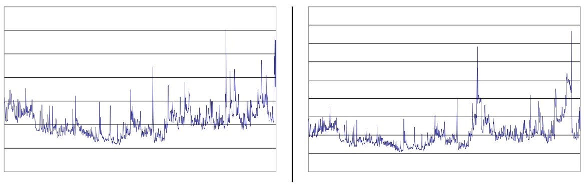

Figure 1: Graphical representation of the vectors of 2 (rather polysemous) terms *exchange* and *profit*

Because of the vagueness related either to polysemy or to lacks of precision (only 873 general concepts), we have to *plunge* our vectors into a specialized semantic space. However, we cannot cut loose from the general ones for two reasons. First, even non-specialized words may turn out to be pivotal in word sense disambiguation of specialized ones. Second, we cannot know beforehand if a given occurrence of a word should be understood in its specialized acception or more a general one.

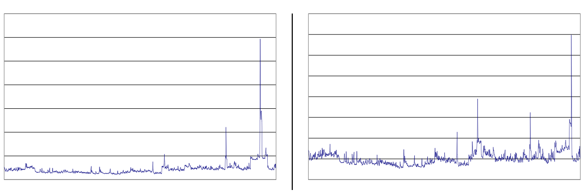

Figure 2: Graphical representation of the vectors of terms *cession* and *benefit*

3 Ontological Extension

The preceding analysis has been done in the space of the general language. We should know evaluate the effect of playing with two spaces: the general one and the general one extended in some localized semantic regions. In all generality, we consider two ontologies G (general) and S (specialized). The space G is supposed to cover roughly the entire semantic space and may describe coarsely any word meaning. The space S is relevant only to words of its own speciality. S is locally much more precise and discriminating than G and the intersection between should not be null. As mentioned before, G and S are generating families for vectorial spaces. In what follows, we would speak of G (resp. S) as the vectorial space defined by the ontology G (resp. S). We would then compare the effect for a given word and their associated meaning vectors to be defined in G or in G ∪ S (aka GS). We would notice that all terms are defined in G, and some of them are defined in GS (the specialized ones).

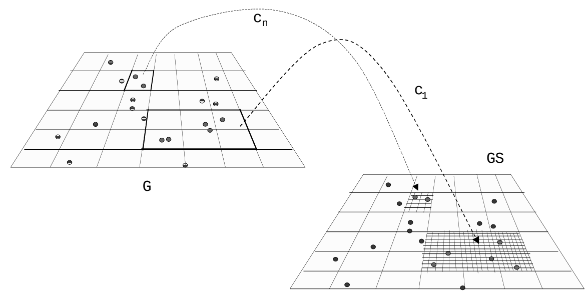

Figure 3: Refinement of the semantic space mesh from G to GS and space correspondences

3.1 Vector Folding and Unfolding

To a concept c_G of G, we can associate a set of concepts of S (called *ontological correspondences*). For instance, the concept of *ECONOMY* of G is associated to the entire subtree of S containing this concept (political economy, market economy, regulated economy, micro-economy, macro-economy, etc.). As a constraint, the set of ontological correspondences should cover S entirely. More precisely, all concepts of S should be linked to G (this is a surjection). However, there may be concepts of G which are not in the semantic field of S and as such are not linked ton any concept of S space. We should not forget here that S in included in G space.

The *Vector Unfolding* function V_U is a projection of a vector v_G of G in the space GS: $v_G \rightarrow v_{GS}$. It allows to refine the vectorial representation if it is relevant to the concepts of S. This operation is also called the *ontological expansion* of v. $V_U\ (v)$ is a vector of $G \cup S$ and is composed of a vector of G followed by a vector of S, and $dim(D(v)) = dim(G) + dim(S)$. The first part of $D(v)$ is v (called Kern). The second part (called Ext) is calculated thanks to the ontological correspondences. The computed vector does not contains any zeros if the vector of the concerned concepts are compact.

The *Vector Folding* function V_F is a projection of a vector v_{GS} of GS over G: $v_{GS} \rightarrow G$. To fold a vector, cutting the extension is sufficient:

$$\begin{aligned} V_F(v)_{GS} &= < x_1, \dots, x_{dim(G)}, \dots, x_{dim(G)+dim(S)} > \\ \rightarrow & \quad < x_1, \dots, x_{dim(G)} > = v_G \end{aligned}$$

By construction, we ensure that links of concepts between G and S are guaranteeing the synchronization between vector components of Kern and the related component in Ext. The folding function is a projection that loose information, in particular if it concerns terms that are both general and specialized (like *exchange* of *market*). However, the activation of the concepts of G reflects the activations of the concepts of S (the Ext part of the vector).

The *ontological refinement* is the learning of a word meaning from its definition in the vector space GS. This should be kept in perspective toward the standard learning of a word meaning in the vector space G. The vector v_{GS} is then constructed both with the general concepts and the specialized ones. The presence in the definition of terms already learning in GS (as they belong to the speciality) would lead to a vector much more relevant than in G. As a consequence, it means that all words of the definitions have been plunged into GS: those which belong to S by refinement, those which don't by unfolding.

3.2 Construction of vectors in GS

Construction of vectors of S. Vectors of S are built exactly like those of G. Basically, the idea is to exploit the ontology organization to construct dim(S) vector of the concepts of S. This procedure takes advantage of the ultrametric distance and the potential transversal activations.

Construction of vectors of GS. The question here is to know how to *add* the vector of G (and which one) to each vector of S that has been produced beforehand. The answers is found by *inverting* the unfolding (inverted unfolding). We apply the very same method than VU, but from a vector of S, we compute a vector of G. We can simply invert the correspondence between G and S in a correspondence list between S and G:

$$\begin{aligned} \mathcal{C} &= < C_G, \{C_{S,1}, \dots, C_{S,n}\} > \\ \rightarrow \quad (< C_{S,1}, \{C_G\} >, \dots, < C_{S,n}, \{CG\} >) &= \mathcal{C}' \end{aligned}$$

The vector of S can be concatenated to the left of its reversed unfolding which produce the G part of the vector of GS:

$$V_U(v_S, \mathcal{C}') + v_S \quad \rightarrow \quad v_{GS}$$

Then, the construction of the vectors of the GS kernel (extensible to all terms of S) and the learning process of the terms on GS can be done as in G.

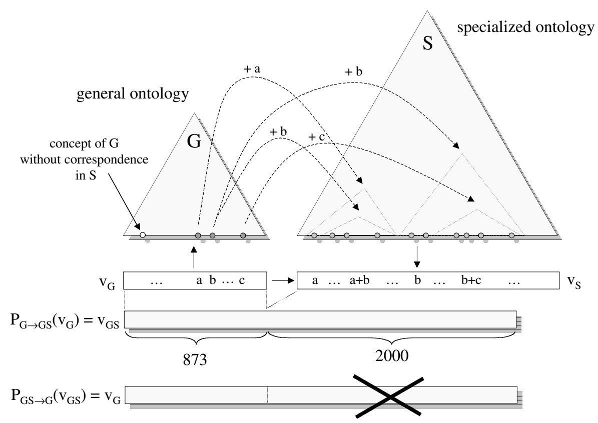

Figure 4: Ontological correspondences and computation the vector extension from the kernel

3.3 Lens Effects

A first property concerns only the composition of Folding and Unfolding functions:

$$V_F(V_U(v)) = v$$

To Unfold and Fold a vector bring nothing and is equivalent to the identity function. But in the general case, the inverse composition does not hold ($V_U(V_F(v)) \neq v$) as we are loosing information in the process. We have also a reduction of the angular distance D_A:

$$D_A(v1, v2) \leq D_A(V_U(v1), V_U(v2))$$

This phenomenon can be understood as the *ontological extension increase the synonymy when focused outside any specific learning.* This can be demonstrated and experimented on most words and word meaning. conversely, the *ontological refinement* can lead to different situations concerning two terms A and B.

1. A *Reduction* of the synonymy between A and B would generally happens. In other words, there is an augmentation of the semantic distance between A and B as they two specialty terms and are more discriminated in GS than in G. The ontological extension acts as a Lens. For instance ‹*public finances*› and ‹*fiscality*› are at a distance of $D_A = 17°$ in G (very close and almost synonymous). In GS, we have $D_A = 69°$.

2. The *Augmentation* of the synonymy by reduction of the polysemy is a more common (and more interesting) event. Polysemy acts here like noise, and the ontological extension acts as a convex lens that would bring together two meanings accidentally separated. For instance, in G both ‹*product*› and ‹*good*› are highly polysemous (with a special mention to ‹*good*› due to the possible multiple morphosyntactic categories). The distance between both word is around 60 in G. In GS, the filtering of the polysemy (for each word, only a subset of the possible meaning finds a realization in S) bring closer associated vectors (and the distance in GS is around 10°).

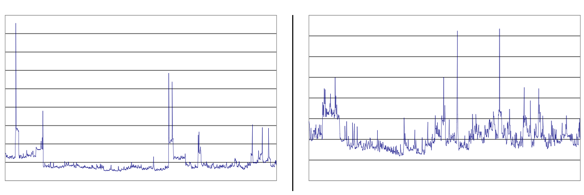

Figure 5: Graphical representation of the vectors of 2 (rather polysemous) terms *product* and *good*

The *Reduction* effect (called also Lens focusing) is an, a priori, expected effect of the ontological augmentation. In our preliminary experiment, we have automatically selected around 200 terms (cf. annex) that are almost synonyms in the general vector space G and are related to the field of economy. By extending the vector representation to GS (by refinement), 95% of them have their distance increased, and are found to be indeed not polysemous in the domain of economy and relevant to this domain. The remaining error is attributed to existing polysemy even in the context of the domain.

The *Augmentation* effect (called also Lens merging) is by itself not as expected than Reduction. In fact, one can make use of a strong analogy with *gravitational illusion* in astronomy. A given observer, may see two images of the same star if there is a heavy gravitational object between them. Here, the effect of polysemy acts as the intermediate object and will make two very different vectors in G for two terms that are quasi synonymous in S. As only the meanings relevant in S will have a vectorial effect for vectors in GS, the terms *under the spotlight of* GS are brought closer. Basically, this situation is common for almost all terms of S that are used everyday in their general meanings (i.e. in G). For instance in French, we had this effect with the following couples (beside ‹*bien*› [good] and ‹*produit*› [product]): ‹*devise*› [motto, slogan, currency] and ‹*monnaie*› [currency, coin, change, mint], ‹*devise*› [motto, slogan, currency] and ‹*monnaie*› [currency, coin, change, mint].

4 Conclusion

This paper has presented a model of thematic representation using the formalism of conceptual vectors and discussed the effect of using specialized domain ontology. The major applications are thematic analysis of texts, construction of large lexical databases and word sense disambiguation. We grounded our research on a computable linguistic theory being tractable with vectors for computational sake. This preliminary work on semantic space extension has also been conducted under the spotlight of ontology and concept set extension, and allowed us to express semantic refinement in terms of conceptual vectors. The experiments we have conducted lead us to the following observations:

When we aim at analyzing, clustering or indexing documents of speciality, the best approach consists in using the union between the general ontology and the ontology of speciality because texts do not only contain technical terms. The shifting between both ontologies has been formalized through the folding and unfolding of vectors.

When the automatic learning of specialized concepts is conducted from the definitions given in dictionaries, this ontology union appears to be very relevant as all words in definitions can contribute activations to the word sense computation in context. In the conceptual vector model, the specialized ontology is much more detailed than the general one contrary to the classical approach of *knowledge tree*. However, we pinpoint the fact that the information quantity to store is smaller in the analysis of technical texts than for general texts. Thus, we have shifted the difficulties related to the size from the ontology to the lexicon.

The speciality ontology allows a finer *mesh* for meaning representation and thus a better semantic discrimination between terms that appear to be semantically close. Inversely, the computation of the meaning and distances on this ontology allows to bring much closer two meanings that are not obviously related in the general ontology. Polysemy, the main characteristic of general lexicon is then circumvent to the benefit of the specialized meaning of the terms.

These conclusions lead us to a more general one: putting in the picture a specialized ontology to process technical text is not only feasible in the vectorial model, but it also allows to unify ontologies, to discriminate word meanings, to confine polysemy and promote the self-emergence of local structuration which could be exploited during the continuous learning process of vectors. Conducted experiments lead to the integration of terminological entries under the form of an ontology of (roughly) 2000 leaf concepts from the OECD and to analyze definitions of a business dictionary (DAFA).

References

C. Barrière and T. Copeck. Building domain knowledge from specialized texts. *Actes de TIA 2001, Nancy*, 2001.

J. Chauché. Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TAL Information*, 1990.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

W. L. Fischer. Äquivalenz und toleranz strukturen in der linguistik zur theory der synonyma. *Max Hüber Verlag, München*, 1973.

G.M. Gwei and E. Foxley. A flexible synonym interface with application examples in cal and help environments. *The Computer Journal*, 1987.

M. A. Hearst. Automated discovery of wordnet relations. *In C Fellbaum ed. "Wordnet An Electronic Lexical Database". MIT Press, Cambridge, MA*, 1998.

T. Hamon and A. Nazarenko. La structuration de terminologie : une nécessaire coopération. *Actes de TIA 2001*, 2001.

M. Lafourcade. Lexical sorting and lexical transfer by conceptual vectors. In *Proceeding of the First International Workshop on MultiMedia Annotation*, Tokyo, January 2001.

Larousse, editor. *Thésaurus Larousse - des idées aux mots, des mots aux idées.* Larousse, 1992.

M. Lafourcade and V. Prince. Synonymy and conceptual vectors. In *proccedings of NLPRS'2001*, Tours, France, July 2001.

E. Morin. *Extraction de liens sémantiques entre termes à partir de corpus techniques.* Université de Nantes, 1999.

OECD. Oecd macrothesaurus - html version - french language. *http://info.uibk.ac.at/info/oecd-macroth/*, 1991.

P. Rodget. *Thesaurus of English Words and Phrases.* Longman, London, 1852.

E. Riloff and J. Shepherd. A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction. *Natural Language Engineering*, 1995.

G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGrawHill, 1983.

S. Verlinde and T. Selva. Dafa - dictionnaire d'apprentissage du français des affaires. *http://www.projetdafa.net*, 2001.

Annex: words related to finance. These French words (or expressions) are the first 200 that have been automatically found to be related to the domain of finance (by comparing their vectors computed in G to those in GS).

$D_A(X,Y)$ in G	*profit*	*benefit*	*finance*	*market*	*product*	*goods*	*good*	*sadness*	*joy*
profit	0	10	19	28	32	31	39	65	39
benefit	25$\nearrow$	0	15	25	34	42	55	67	38
finance	56$\nearrow$	41$\nearrow$	0	17	27	40	45	68	65
market	62$\nearrow$	48$\nearrow$	48$\nearrow$	0	20	21	43	67	55
product	68$\nearrow$	31$\nearrow$	35$\nearrow$	59$\nearrow$	0	11	19	59	45
goods	59$\nearrow$	63$\nearrow$	42$\nearrow$	48$\nearrow$	25$\nearrow$	0	25	57	41
good	65$\nearrow$	52$\searrow$	62$\nearrow$	39$\searrow$	36$\nearrow$	12$\searrow$	0	32	25
sadness	68$\nearrow$	62$\searrow$	71$\searrow$	74$\nearrow$	79$\nearrow$	65$\nearrow$	56$\nearrow$	0	28
joy	55$\nearrow$	42$\nearrow$	67$\nearrow$	72$\nearrow$	59$\nearrow$	52$\nearrow$	36$\nearrow$	22$\searrow$	0

Figure 6: Table of distance between some word in G and in GS

Some of them are very polysemous in the everyday language (these items are stared).

non-affectation, mandatement, cambiaire, quasi-monnaie, réescompter, ordonnancer*, réescompte, financier, cash-flow, eurodollar, intérêt, escompte, arbitrer, crédit-bail, mandater, euro-émission, script, arbitragiste, caméral, versement, transférabilité, convertibilité, rencaisser, euromarché, monétaire, réévaluation, surcapitalisation, consolidé, inconvertibilité, titrisation, budgétisation, autofinancement, bancaire, euromonnaie, finance, holding, ordonnancement, obligataire, dilutif, grand-livre, monométallisme, scripturale, recouvrable, compte chèques, référé, report*, impasse*, désaffectation, désaffecter, ordonnateur, exigibilité, reporteur, tenue*, fiscalisation, bas de laine, convertissement, banquier, déplafonner, revenant-bon, gestion, reversement, dépense, cambiste, eurobanque, déplafonnement, créditer, perceptible*, certificat*, coupon*, agio, euro-obligation, percevable, quotité, fiscaliser, open market, liquidité, économiser, maltôtier, changeur, épargne, C.C.P, rehaussement, swap, avoir*, imputer, encadrement*, contre-valeur, geler*, arbitrage, découvert, monétarisme, devise*, publicain, lucre, débirentier, autofinancer, eurodevise, conversion*, foirail, assurance-crédit, décaissement, défiscaliser, accréditif,, comptabilité, cogestion, décote, compte*, scriptural, escompteur, inconversible, inconvertissable, commanditer, kip, encaisse, bulletin de paie, pécuniaire, agent de change, profit*, revolving, avaliser*, gestionnaire, fictif*, autogestion, escompter, monnaie*, endossataire, dépréder, crédit, remettant, peseta, investisseur, date*, cofinancement, indexer*, casquer, monométalliste, moins-value, butoir, rémunération, en-cours, ratio, fiscal, positionneuse, réfaction, guelte, régulariser, évaluation, comptant, libératoire, perception*, créditeur, positionniste, emprunt, rétribution, dégeler*, numéraire, argentier, coté, économe, participatif, endosser*, se faire un matelas, capitalisation, épargnant, arrêté, endosseur, admission*, faux frais, paye, régie, immobiliser*, roulement*, remboursement, pension*, paie, levier*, affidavit, loyer, trésorerie, finances, vatu, chéquier, change, budgétaire, compartiment, Sicomi, caisse de dépôts, percepteur, caissier, accréditer, appointements, sou, endossement, rémunérateur, soudoiement, achat, inchangeable, conglomérat, limiter*, vénal*, bazar, établissement*, précompte, billetage, ressources, fixage, écu*, réemploi, cellérier, monétiser, penny, contrepartiste, monnayer, crédencier, sous-traitance, endos, boursicoter, coteur, payeur, trafiquer*, économique, reverser, décaisser, certifier*, somptuaire, traitant, questure, établissement de crédit, bazardage, peseur*, contribuable, multinationale, cagnotte, coter, eurofranc, faire sa pelote, bureaucratique, affranchissement, subvention, emprunteur, payer*, amortissable, douane*, dépocher, capitaliser, boursier, fermage, vente, syndic, compétitif, dégrèvement, kolkhoz, domicilier, microéconomie, douiller, piastre, C.F.A, acquit-à-caution, échange, souk, cotable, remisier, contrepartie, trust, vente aux chandelles, frappage.