# Making people play for Lexical Acquisition with the JeuxDeMots prototype

Mathieu Lafourcade

# Making people play for Lexical Acquisition with the JeuxDeMots prototype

Mathieu Lafourcade
LIRMM (CNRS - Université Montpellier 2) - FRANCE
e-mail : Mathieu.lafourcade@lirmm.fr

## Abstract

Lexical and semantic information are difficult to collect either automatically or manually, and as being instrumental in many (if not all) NLP applications, there is a serious bottleneck for advances in this research field. In this paper, we propose a novel approach by making people play some king of associative word games. Doing so, we can memorize associations where a pair of players agrees. We present the principles of the game and some insights of the data collected so far.

## 1 Introduction

Many NLP task (such as WSD, PP attachment, conjunction scope, Anaphora resolution, Information Retrieval, etc.) need lexical and semantic information usually found in thesaurus or ontologies. Such resources can be produced manually (like Roget or Wordnet) or automatically from text corpora (Spark Jones, Grefenstette, Lin). Broadly speaking manually built resources tend to be semantically found, and automatically produced one are more based on word distribution.

Furthermore, we need lexical information of various nature, of course semantic (like on ontologies) with hypernym/hyponym relations, holonym/meronym (part-of, whole) but also more on usage like lexical functions (Meltchuk) and mental lexicon which usually stress on free term associations.

Sadly, manual production is slow and costly and event its quality can be put at stake. Automatic construction highly depends on the chosen corpora and its accurateness strongly decreases as more subtle relations between words are to be extracted from texts.

Another apparently unrelated fact, (von Ahn, 2006) pointed out by that during 2003 in the United-Stated a cumulative 9 billions hours have been spent by people playing Solitaire.

If we put these two observations together, we can ask ourselves if it could be possible to make people constructing lexical data by playing games. The answer is certainly positive, but the difficult question is how to organize such games in a way that ensure both quality and coverage in an acceptable timeframe.

This research is part of the Jeux de Mots and Papillon projects. The "JeuxDeMots" prototype can be played at the following address: http://www.lirmm.fr/jeuxdemots

## 2 A lexical graph

The lexical data we aim at is strongly based on the notion of graph, reminiscent of lexical network. All data is represented through nodes and relations between nodes. (Polguère, 2006)

### 2.1 Objects

Objects can be nodes or relations. Although of different usage, they have roughly the same structure, as follows:

- A name/id
- A weight
- A confidence level
- A creation date

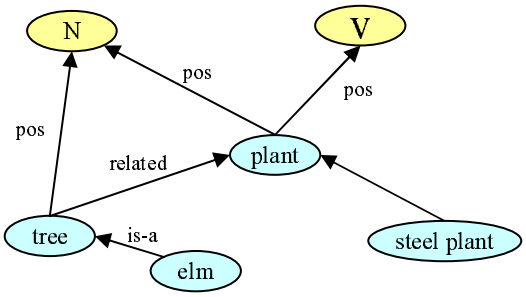
The name/id is the basic information either under the form of a string or a numeral and his unique. The weight refers to the number of time this object has been activated. The confidence level is an evaluation of the validity of this

object. The creation date permits to compare object as been ancient or recent.

Generally speaking old and "heavy" object have a high confidence level, but it may happen to have ancient or moderately heavy object with low confidence level, for example common errors.

### 2.2 Nodes

The most basic type of node contains a word form, which is generally a term (usually a lemma) and refers to one or several meanings in standard dictionaries.




We have also POS (or other morphological information) nodes. A term node is related to an occurrence of POS node instead of having it directly encoded in the node structure. Thus approach is generic and can handle any type of information that may be uncertain and fuzzy (at least during the data construction).

### 2.3 Relations

We have different type of nodes, such as terms, acceptions, usage nodes but also POS nodes. Those nodes are linked by different types of relations, such as classical lexical relations:

- Synonym
- Antonyms
- Locutions/collocations
- Derivations
-

Classical ontological relations are also present:

- Freely associated ideas
- Hypernym (is-a) and Hyponym
- Part of and Whole
- Typical agent / patient / instrument (for predicates)

Few lexical functions (in the Meltchuk spirit) like *magn* and *antimagn* are also represented through relations.

The weight relates to the number of time this relation (or node) has been encountered through people games. The confidence value is a measure of the validity of this relation (or node) and is related to the 'efficiency' of the player (see below). We do not foster this paper on confidence, which can be accessed through "riddle" games (which are the "inverted" approaches compared to association games)

## 3 Game Model

JeuxDeMots is a two player blind game based on agreement. Blind implies that the other player is unknown until the end. Agreement means that both players will always score the same amount of points according to what they have proposed in common.

### 3.1 Principle

At the beginning of a game session the player is given an instruction related to a target term? For example: *give any term that is related to "cat"*.

Then the user has a limited amount of time (around one minute), to enter as many propositions as possible.

At the end of the allowed time, player A proposals are compared to those of player B. points are earned on the basis on the common proposals (that is to say, the intersection).

This game model has been used in other context, like in the image indexer of Google (von Ahn, 2003). But as far as we know, no experiment has been undertaken for lexical domains.

Terms in agreement are added to the lexical network with a relation corresponding of the game instruction linking to the target word. If the relation already exists, its weight is increased.

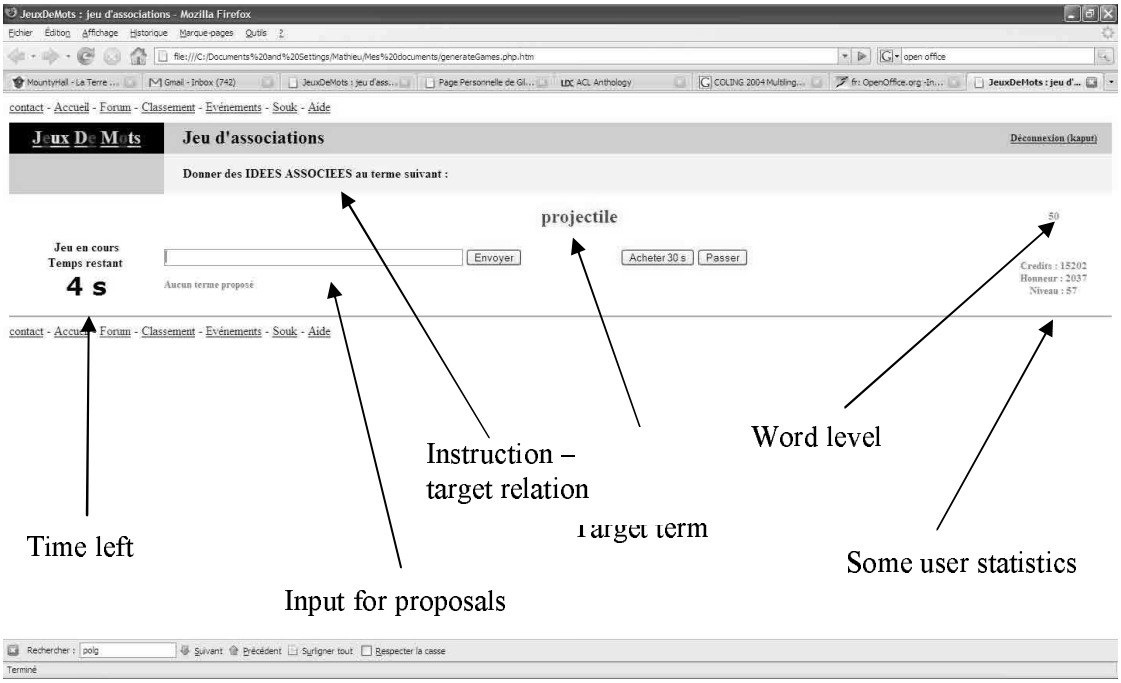



**Figure 1**. A typical Jeuxdemots game presentation

### 3.2 Asynchronous and Symmetric Play

Players A and B are not playing at the same time, for several reasons that range from interaction modalities to technical ones.

A game session is randomly chosen between two types. Either, the player will make the first half of a game (we say the player is creating a game), either the player will make the second half of a game previously created by another player (we say the player is concluding a game).

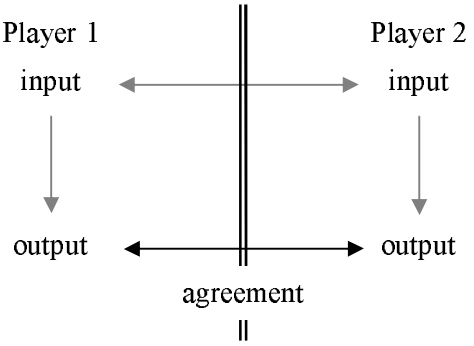
When concluding a game, the player is given his/her score immediately, and the game creator (the player who has created this game) will be notified by email of the score.

### 3.3 Mutual agreement

The scoring mechanism is based only on the terms where the two players agreed. The more terms they will have in common, the more points they might gain.

However, for a given term, the more it has been already proposed, the less it will be rewarding. In the network model this translated simply by rewarding in inverse proportion of the weight this relation occurrence already has.

The relation between the target term and the proposed term is then increased. If an agreed term is not in the network, it is added. Later on, it can be proposed as a target term.




### 3.4 Taboo mechanism

When a relation occurrence reaches a given threshold, the term is budded as "taboo" and will not give any more points. Taboo words are displayed to the player during the game, encouraging him to make other proposal. This mechanism is instrumental in augmenting the coverage in the lexical network.

For instance, for the "associated idea" relation and the word "cat", terms like "feline", "animal" and "mammal" become very quickly taboo. Players have then to figure out other less straightforward terms that could be associated to "cat"

As we keep the creation date of relations and the date it became taboo, we can order taboo relations event their score are similar.

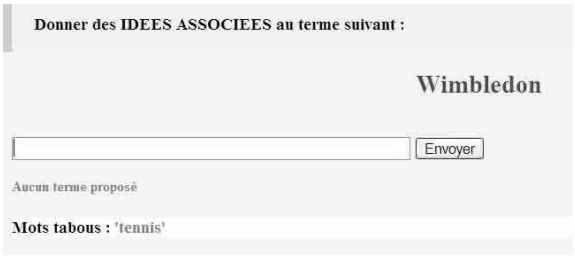



**Figure 2.** taboo word (tennis) for Wimbledon.

In figure 2, for the target word "Wimbledon", the term "tennis" has become taboo, inducing players to find other possible associations.

### 3.5 Target selection

The more specific the relation, the more difficult it is to propose a target word that would provide interesting (or even existing) associated terms. For instance, for the relation "part of" it would a bit odd to propose the verb "to eat" as a target.

**POS** The first simple strategy if to select target words according to their part of speech. For instance, for relations related to predicates, we only select term dubbed as Verbs.

Potential nodes There is a special type of node called "Potential node". For instance, there is one occurrence of the "Is-a Potential Node". If a given term is found to be related (through is-a) to other terms, then it will be expected to have a high potential for this relation, and be linked to the potential node with a strong weight. This weight will increase or decrease according to the game linking activity. The term selection for a game is directly related to its potential. That way, we can propose more and more meaningful term for a given relation. Also, when a term/relation pair has many taboo words as depleted possible associations, the potentiality eventually decrease, making the term less selected for games.

## 4 Making the game addictive

Although, these aspects are slightly off the topic of this paper, it is noteworthy to discuss some aspects that make the game addictive. We present as a broad picture few aspects of player satisfaction and what make people come back playing.

### 4.1 Satisfaction from Agreement

It appears that people are delighted when they match to each other, especially when the term in common where not straightforward, or the target term was difficult. Interesting enough some people asked us if it would be possible to get some other people email to contact them.

### 4.2 Satisfaction from ranking

There is a ranking among players which also fosters satisfaction or develop a competitive feeling especially amongst top players. Beside credit points which allow the player to "buy" various item, such as time, during a game, there are "honour points. Honour points basically measure the quality of the contribution of the player to the database. The more efficient a player, the more impact he/she have of the increase of weight for relation and nodes. This aspect tempers the impact of spammers, i.e. players that play a lot but without many proposals. Also, the possible impact of bots which tend to make quantity over quality is strongly reduced.

### 4.3 Satisfaction from progression

Upon registering to the game, a player will only play though associated ideas. While progressing in points, the player will be offered to "buy" some other relations to play with. This mechanism allows players to get used to the game with "easy" relation before playing with harder ones. Moreover, the feeling of progressing in competences has a strong motivation effect to people.

### 4.4 Coming back: short play

The typical game is around one minute. As such, the idea of playing few games is not a deterrent for someone who wants to makes a break during work. Although, we measured that the average playing session is around 25 minutes, which correspond generally to many more games than anticipated by the player.

### 4.5 Coming back: mailing

When a game is concluded by a player, the game creator (the person who inited the game) is notified by email. This tends to induce people to come back.

## 5 Experiments

In one month time since the launching of the first beta version of JeuxDeMot (beginning of July 2007), around one hundred persons registered to the game. No special advertisement has been made, thus people went to JeuxDeMot only by word of month. Players have produced around 20000 relation occurrences. Around 15000 of them are of "associated idea" type, 2000 for synonyms, 2000 for "is-a" and other are distributed among other relations.

Currently, we set up JeuxDeMots only for French but the adaptation to other language would not be difficult. We bootstrapped the game with a word list of 150000 terms. More than 2000 new terms (mostly compound words or proper names) have been "discovered" since then.

We compared the data obtained in JeuxDeMots (JDM thereafter) so far with those of Euro Wordnet French (EWF thereafter). Quantitatively, there are 151614 terms in JDM and 23066 terms in EWF. They are 157 terms of EWF not in JDM, mostly specific compound terms like "communiquer par réseau", "phénomène chimique" that are useful for the concept hierarchy. We collected so far 21807 relations in JDM compared to more than 100000 in EWF.

From a qualitative point of view, we compared a sample of 100 very common terms (actually the first 100 with the highest weight in JDM). If we just consider hyperonyms and hyponyms, around 35% on the hypernym in EWF were inexistent in JDM. On the other hand, several hypernyms are generally present in JDM (not only because of polysemy but because of word usage) and only 80% of them are inexistent in EWF. Il 97% of the case the associated were at least correct. The remaining 3% were "common errors" of confusions (which by themselves are interesting data). The data collected of JDM bring of a lot of novelty but much less precision than hand crafted data like EWF.

## 6 Some Perspectives

People mostly play with the most generic relation at the beginning, but progress with more and more complex and rewarding ones. Analysis of the lexical network can lead to automatic discovery of word meanings, to be validated by players through games.

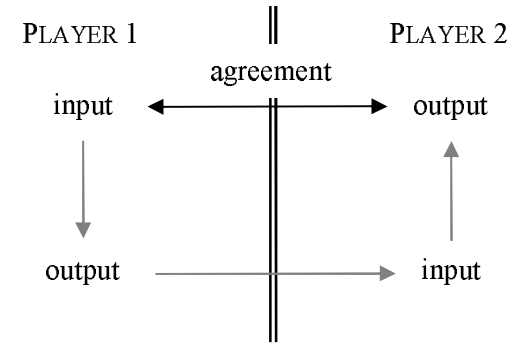
### 6.1 From word to word usage

Slow construction of ontological information from simple to complex relations, seems to be a good approach for discovering word usages (if not acceptions). As it seems difficult to ask player to build definition new approaches should be considered. Taking sentences containing the target word can be a good way to give a context to the player for selecting word meanings for polysemous terms.

### 6.2 More game types

The unique game type implemented at present is of associative type. That is to say, for a couple (relation type, target), the task at hand is to give related terms. This type of game becomes unsuitable for very specific relation type as they are lexically very sparse. In the same framework, we are modelling and testing some different game types, one on which is a "guessing game". From a target word, the player would be able to select relation template to be completed. The goal is to make the other player guess what could be the target word.

When completing a game, the player will be displayed the different completed template during time and able to propose possible candidates. The games finished when the player has found the target term or time runs out. Both players will be rewarded the same way in inverse proportion to the time taken to find out the target term Let's give a very simple example. The task would be the other player to guess about the target word "milk". The player can select the template "made by" and complete it as "made by cow". The game model is then rather different as being asymmetric, but still based on agreement and blind.



## 7 Conclusion

In this paper, we have presented some of the principle of JeuxDeMots, a web-based game with the purpose of building a lexical network. Although this experiment is quite recent, we are confident that if the game attracts people, we will eventually obtain lexical data with quite a large coverage and high quality. Still some in-depth and systematic comparisons are still needed for existing manually build similar resources, like Wordnet, but to our knowledge there is no large scale data on associated ideas.


### Acknowledgements

I would to thank all the people who contributed by their ideas to the JeuxDeMots prototype, among others : G. Sérasset, M.Y Monod, and A. Joubert.


### References


von Ahn, L., and Dabbish, L. Labeling Images with a Computer Game. In ACM Conference on Human Factors in Computing Systems (CHI), 2004, pages 319-326.

von Ahn, L., *Humain Computation*, CMU, Slide presentation., 2006.

Dongyiang Y., Powers, David M. W. . Verb *Similarity on the Taxonomy of WordNet.*

Olivier Ferret; Michael Zock Enhancing Electronic Dictionaries with an Index Based on Associations. ACL 2006

Jeux de Mots, http//gohan.imag.fr/jdm, 2005

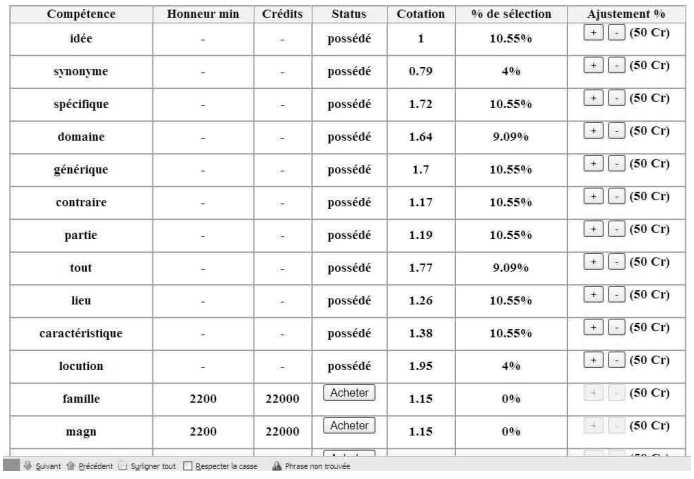
Kilgariff, A. *Thesauruses for Natural Language Processing*. Keynote lecture. Proc. Natural Language Processing and Knowledge Engineering (NLPKE). Beijing, October, 2003.

Fellbaum, *WordNet An Electronic Lexical Database,* Edited by Christiane Fellbaum, MIT Press, ISBN-10: 0-262-06197-X, 1998.

Polguère, A. *Structural Properties of Lexical Systems: Monolingual and Multilingual Perspectives,* Proceedings of the Workshop on Multilingual Language Resources and Interoperability, COLING 2007.

Rada F. MIhalcea, Moldovan D. I., A highly accurate bootsrapping algorithm for word sense disambiguation, IJAIT, Vol 10, No 1-2 2001.

Resnik P., Semantic Similarity in a Taxonomy: An information-Based Measure and its Application to Problems of Ambiguity in natural Language., in JAIR, 1992

| Compétence | Honneur min | Crédits | Status | Cotation | % de sélection | Ajustement % |
|---|---|---|---|---|---|---|
| idée | - | - | possédé | 1 | 10.55% | + - (50 Cr) |
| synonyme | - | - | possédé | 0.79 | 4% | + - (50 Cr) |
| spécifique | - | - | possédé | 1.72 | 10.55% | + - (50 Cr) |
| domaine | - | - | possédé | 1.64 | 9.09% | + - (50 Cr) |
| générique | - | - | possédé | 1.7 | 10.55% | + - (50 Cr) |
| contraire | - | - | possédé | 1.17 | 10.55% | + - (50 Cr) |
| partie | - | - | possédé | 1.19 | 10.55% | + - (50 Cr) |
| tout | - | - | possédé | 1.77 | 9.09% | + - (50 Cr) |
| lieu | - | - | possédé | 1.26 | 10.55% | + - (50 Cr) |
| caractéristique | - | - | possédé | 1.38 | 10.55% | + - (50 Cr) |
| locution | - | - | possédé | 1.95 | 4% | + - (50 Cr) |
| famille | 2200 | 22000 | Acheter | 1.15 | 0% | + - (50 Cr) |
| magn | 2200 | 22000 | Acheter | 1.15 | 0% | + - (50 Cr) |

Suivant Précédent Surligner tout Respecter la casse Phrase non trouvée

**Figure 3.** The table for competence (relation type) buying. The "Cotation" display the value of the relation, thus that the player can select the more profitable in terms of points.

**Relations collected for the term** « chat »

chat `---r_isa:220-->` animal
chat `---r_associated:190-->` félin
chat `---r_isa:190-->` félin
chat `---r_associated:190-->` animal
chat `---r_associated:130-->` minou
chat `---r_associated:130-->` chien
chat `---r_associated:110-->` chatte
chat `---r_has_part:110-->` patte

chat ---r_isa:110--> mammifère
chat ---r_has_part:100--> oreille
chat ---r_has_part:100--> poil
chat ---r_associated:100--> souris
chat ---r_has_part:80--> griffe
chat ---r_isa:80--> animal de compagnie
chat ---r_associated:80--> griffe
chat ---r_associated:80--> minet
chat ---r_associated:70--> poil
chat ---r_has_part:70--> yeux
chat ---r_associated:70--> miauler
chat ---r_associated:70--> moustache
chat ---r_associated:70--> ronronner
chat ---r_associated:70--> chaton
chat ---r_associated:70--> miaou
chat ---r_has_part:70--> queue
chat ---r_associated:70--> matou
chat ---r_has_part:60--> coussinet
chat ---r_associated:60--> félidé
chat ---r_has_part:60--> oeil
chat ---r_has_part:60--> pattes
chat ---r_associated:60--> siamois
chat ---r_has_part:60--> langue
chat ---r_has_part:50--> tête
chat ---r_has_part:50--> griffes
chat ---r_associated:50--> griffes
chat ---r_associated:50--> litière
chat ---r_isa:50--> carnivore
chat ---r_associated:50--> sieste
chat ---r_isa:50--> compagnon
chat ---r_associated:50--> croquette
chat ---r_associated:50--> miaulement
chat ---r_associated:50--> zoologie
chat ---r_associated:50--> queue
chat ---r_has_part:50--> moustache
chat ---r_has_part:50--> poils
chat ---r_isa:50--> félidé
chat ---r_associated:50--> persan
chat ---r_associated:50--> doux
chat ---r_isa:50--> être vivant

chien ---r_associated:270--> chat
félin ---r_associated:240--> chat
minou ---r_associated:200--> chat
chatte ---r_associated:120--> chat
patte ---r_associated:120--> chat
souris ---r_associated:110--> chat
animal de compagnie ---r_hypo:100--> chat
animal ---r_associated:70--> chat
poil ---r_associated:60--> chat
gris ---r_associated:60--> chat
Félix Ier ---r_associated:60--> chat
chaton ---r_associated:60--> chat
miaou ---r_associated:60--> chat
lynx ---r_associated:60--> chat
Egypte ---r_associated:50--> chat
souris ---r_isa:50--> chat
caracal ---r_associated:50--> chat
pattes ---r_holo:50--> chat
aiguille ---r_associated:50--> chat
pattes ---r_associated:50--> chat
chien ---r_domain:50--> chat
griffe ---r_associated:50--> chat

**100 most frequent term in JeuxDeMots after 3 months**

| | | | |
|---|---|---|---|
| voiture -- 504 | oiseau -- 276 | yeux -- 224 | chien -- 184 |
| maladie -- 495 | politique -- 274 | religion -- 223 | prénom -- 182 |
| mer -- 484 | chat -- 272 | art -- 222 | France -- 182 |
| musique -- 444 | homme -- 270 | métal -- 216 | chanteur -- 180 |
| animal -- 439 | froid -- 269 | médecine -- 214 | cheval -- 180 |
| livre -- 432 | arbre -- 268 | vélo -- 208 | bleu -- 178 |
| eau -- 423 | soleil -- 268 | pluie -- 206 | temps -- 176 |
| guerre -- 417 | cinéma -- 262 | neige -- 206 | père -- 174 |
| Harry Potter -- 382 | peinture -- 258 | vin -- 205 | acteur -- 174 |
| sport -- 372 | blanc -- 258 | porte -- 204 | ciel -- 170 |
| avion -- 365 | école -- 254 | malade -- 202 | mort -- 170 |
| magie -- 362 | fleur -- 244 | dormir -- 200 | médicament -- 170 |
| argent -- 356 | médecin -- 244 | lit -- 196 | température -- 169 |
| ville -- 348 | langue -- 242 | corps -- 194 | famille -- 168 |
| pays -- 338 | jeu -- 240 | enfant -- 194 | ballon -- 166 |
| train -- 336 | tennis -- 240 | préservatif -- 194 | moteur -- 166 |
| sexe -- 330 | ordinateur -- 236 | informatique -- 194 | plage -- 166 |
| bateau -- 318 | noir -- 235 | fruit -- 193 | espace -- 166 |
| sorcier -- 310 | plante -- 234 | peintre -- 191 | machin -- 164 |
| amour -- 302 | tête -- 234 | course -- 191 | nuit -- 164 |
| femme -- 302 | rouge -- 233 | internet -- 188 | téléphone -- 164 |
| couleur -- 298 | bois -- 232 | pied -- 186 | verre -- 164 |
| maison -- 298 | manger -- 228 | baguette -- 184 | félin -- 162 |
| président -- 294 | roman -- 227 | écrivain -- 184 | rugby – 160 |
| film -- 291 | feu -- 226 | chaud -- 184 | |