# Propa-L: a Semantic Filtering Service from a Lexical Network Created using Games With A Purpose

Mathieu Lafourcade, Karën Fort

# Propa-L: a Semantic Filtering Service from a Lexical Network Created using Games With A Purpose

**Mathieu Lafourcade***, **Karën Fort**†

* LIRMM (CNRS - Université Montpellier 2), Montpellier, France, mathieu.lafourcade@lirmm.fr

† Université de Lorraine/LORIA, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy, France, karen.fort@loria.fr

**Abstract**

This article presents Propa-L, a freely accessible Web service that allows to semantically filter a lexical network. The language resources behind the service are dynamic and created through Games With A Purpose. We show an example of application of this service: the generation of a list of keywords for parental filtering on the Web, but many others can be envisaged. Moreover, the propagation algorithm we present here can be applied to any lexical network, in any language.

**Keywords:** game with a purpose, crowdsourcing, lexical network, semantic filtering.

## 1. Introduction

There are many applications that need lists of semantically-related terms. For example, most parental control systems are based on so-called *white* (authorized) and *black* (unauthorized) lists. This raises two issues: $(i)$ the cost of creation of these lists and $(ii)$ the rapid obsolescence of the created lists. Thus, for parental control systems, according to (Cioffi et al., 2009): "it is really difficult to maintain updated the lists because of the high number of new resources daily introduced".

This can be seen as a more general task of producing lists of semantically-related terms. This has been studied, in particular in (Habert et al., 1998), using syntactic similarities or similarity methods using WordNet (Fellbaum, 1998), as in (Varelas et al., 2005). However, these methods do not allow for the creation of dynamic resources (resources that evolve in time) and they involve resources that were costly to develop.

The approach we propose allows for the creation of dynamic resources for a limited price. First, relying on crowdsourcing leads to the creation of lexical resources where terms associations and meanings can be updated almost in real time with real speakers usage. Second, term activation by propagation of information on a lexical network can be combined with filtering constraints. The higher the activation, the higher the probability that the term belong to the set defined by intention (constraints plus one or several seed terms). This type of approach allows for the computation of a term list that can be available on the spot.

We first present the lexical network and the Games With A Purpose that originated the resource, then we detail the filtering service and show the obtained results on a parental control example.

## 2. Lexical Network and GWAPs

### 2.1. JeuxDeMots, an Associative Game

At the heart of the system is the semantic network built through JeuxDeMots (Lafourcade, 2007), a Game With A Purpose (GWAP, see (von Ahn and Dabbish, 2008), (Siorpaes and Hepp, 2008) and (Thaler et al., 2011)) that has been on-line for seven years now, for its French version. Many lexical-semantic networks have been manually developed, like WordNet (Fellbaum, 1998), and its multilingual version (Vossen, 1998), or adapted as WOLF (Sagot and Fišer, 2008), among others, like HowNet (Lenat, 1995) and (Dong and Dong, 2006). Alternative similar resources have been developed by automatic extraction and crossing, like for instance BabelNet (Navigli and Ponzetto, 2012) from Wikipedia. In the case of JeuxDeMots, the resource is built through the gaming activities of a large number of players.

The principle behind JeuxDeMots is quite straightforward: players are asked to enter ideas associated with a term chosen by the system (see figure 1). Players get points if they have answers in common with other players. The more original the answer, the more points are rewarded, but the higher the risk of having no intersection at all with others.


Figure 1: JeuxDeMots: give ideas associated with "manger" (to eat).

Figure 2 shows the summary (recap) of a game. The player scores points: honor points that count for the ranking and credit points that allow to gain some control on the game. Having some experience, players can get to play games on specific lexico-semantic relations, like *is-a*, *hyponym*, *characteristic*, *location*, *agent*, *patient*, etc. Terms in common for a specific game are added (or reinforced) in the lexical network. The mechanism behind the game induces that people performing well will help constructing a quality resource and players below average will have a negligible im-

pact on the resource contents.


Figure 2: JeuxDeMots: recap of a game involving "casseur" (hooligan).

A galaxy of pseudo games (the definition is given in section 2.2.) were created around JeuxDeMots, including mainly LikeIt[1], AskIt[2] or ColorIt[3], that add information to the network. SexIt is one of them.
The created resource is freely available under a Creative Commons License BY-SA 2.0 FR.[4] Note that there is only one lexical resource that is created and enriched , while each game focuses on different modes of acquisition or specific information.

### 2.2. SexIt, an Opinion Game

SexIt[5] allows players to identify a term as being related to sex or not (see an example in figure 3, for "beach"). It is a *pseudo game*: there is no score, no ranking, no gain, only a poll-like comparison with other players. It is played only because the topic is fun: we would have had more difficulties collecting data on taxation, for example. This is an important limitation for language resource building using that kind of interaction. Collecting information about vocabulary related to sex can be necessary, for example to detect pornographic contents. However, such information is insufficient for discriminating pornography from courses in biology, reproduction or anatomy which, while not being pornographic, contain their share of sex-related terms.


Figure 3: SexIt: is "beach" (*plage*) related to sex?

[1] http://www.jeuxdemots.org/likeit.php
[2] http://www.jeuxdemots.org/askit.php
[3] http://www.jeuxdemots.org/colorit.php?thema=-1
[4] http://www2.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR/
[5] http://www.jeuxdemots.org/sexit.php?shy_yes=1

SexIt is built on the semantic network obtained from JeuxDeMots. It enriches the network by adding links to a *Sex* and a *NoSex* nodes. The links are weighted according to the number of players who chose the relationship. As of October 2013, SexIt allowed to process approximately 11,000 terms – from "peau" (skin) to "à la plancha" (plancha-grilled) or "Monica Lewinsky"–, and generated 52,000 hits.
The main issue with SexIt is that the positive responses cover a large number of topics, including reproduction, anatomy, contraception, pornography, seduction, ... If we want to use the resource in a parental filtering application, for example, we need to filter it out.
We therefore developed a service, allowing both for the filtering of the network, according to various criteria, and for the selection of terms by signal propagation from a point of view.

## 3. Propa-L, a Filtering Service

### 3.1. Algorithm

A *filter* is a set of conditions that a term should verify. In the context of the lexical network, a condition is the existence of the specific relation between the tested term and another term. For example, the filter $\{x \xrightarrow{has-parts} wings\}$ specifies that the term x should have wings. Conditions of the set can be conjunctive (*and*) or disjunctive (*or*). For example, the filter $\{x \xrightarrow{has-parts} wings \ \& \ x \xrightarrow{carac} fast \ \& \ x \xrightarrow{is-a} vehicle\}$ specifies that considered terms should have wings, be fast and being a vehicle.
Selecting all terms that have at least one vote in being related to sex can be done with the filter $\{x \xrightarrow{informations} SEX\text{-}YES\}$. More complex conditions can be devised, like comparing the strength of several relations. During the process, the conditions are evaluated on each considered term and memorized.
The *propagation* algorithm is presented below (see Algorithm 1) and consists first in choosing a starting term along with a set of filtered terms and then propagating a large number of signals along the lexical network. A signal sent from the starting term with a strength $k$, will randomly wander in the lexical network considering only filtered terms. The selection of the next term is done pseudo-randomly with a probability proportional to the strength of the relations. The stronger a relation, the higher the chance for it to be selected. Each time a signal reaches a term, the term value is increased by one, and the strength of the signal decreased by one. Hence, a signal with strength 3 can wander up to distance 3 from the starting term. A simplified view of the JeuxDeMots network, before and after propagation, is presented in figures 4 and 5, respectively.

### 3.2. Results

Propa-L is freely available online.[6] The interface provides the user with a number of parameters to fill in:

- a starting term: this term is the focus from which other terms of the lexical network are going to be enumerated or skipped;

[6] http://www.jeuxdemots.org/propagate.php

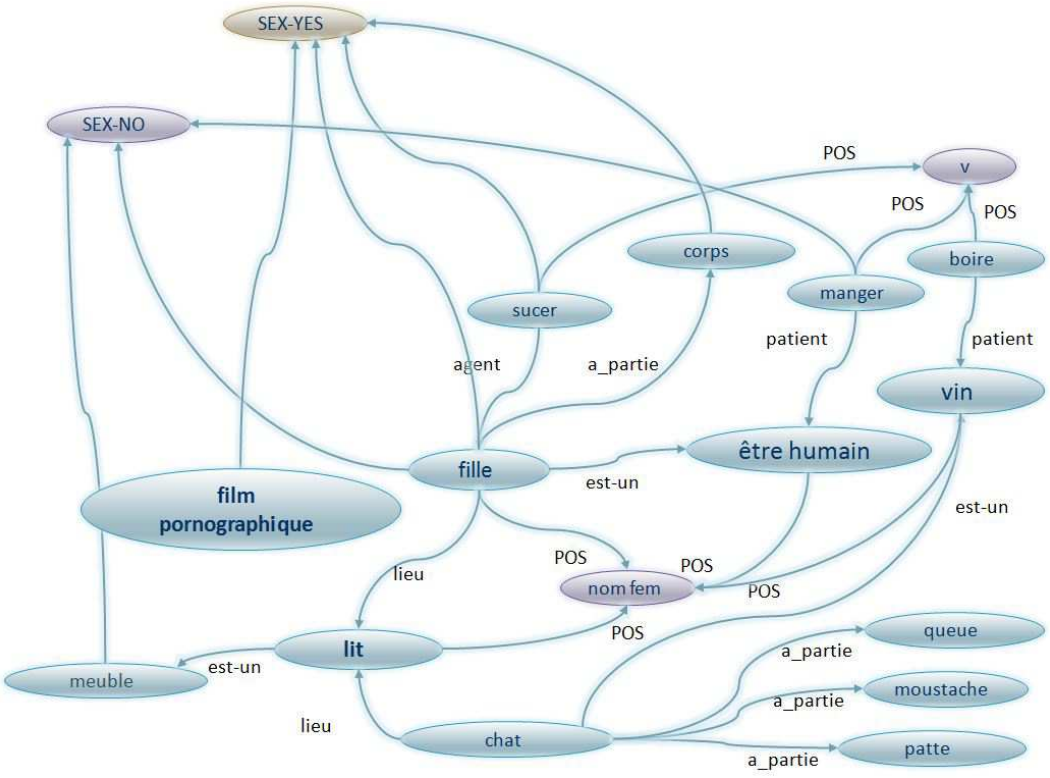


Figure 4: An overly simplified view of the `JeuxDeMots` lexical network: words and word meanings are linked with typed and weighted relations. Any type of information take the form of a node linked to some terms. In the case of `SexIt`, two internal nodes, SEX-YES and SEX-NO, are pointed by terms when users are playing. This principle can be generalized to any type of information polling.

**Data**: A lexical network $L$, a starting term $S \in L$, a signal strength $K$, a number of signals $n$, a filter $F$.
**Result**: A modified version of $L$ where node weights are set.
initialize all node weights to 0;
**while** *n > 0* **do**
  current_node ← S;
  k ← K;
  **while** *k > 0* **do**
    current_node ← random_next_node(current_node);
    weight(current_node) ← weight(current_node) + 1;
    **if** *not(current_node ≺ F)* **then**
      k ← 0;
    **else**
      k ← k-1;
  n ← n-1;

**Algorithm 1**: The propagation algorithm with filtering. If a term reached by the signal has been filtered out, then the signal is blocked. The selection of a next node is done pseudo randomly among the neighbors of the current node.

- a number of signals to be sent (cycles);
- the strength of the signal;
- a filter (can be empty).

The request produces a weighted list of terms. For example, with the starting term *biologie* (biology), 100,000 signals of strength 3 and a filter consisting in keeping only terms having at least one positive link in `SexIt`, we obtain the following list:

• science 13574 • cellule 6733 • génétique 5876 • vie 5609 • biologie 5005 • vivant 4114 • anatomie 3904 • médecine 3717 • zoologie 3002 • ADN 2483 • animal 2404 • animaux 2054 • physique 2035 • gène 1968 • reproduction 1890 • os 1794 • santé 1586 • gynécologie 1435 • botanique 1410 • chromosome 1358 • matière 1342 • sexe 1303 • orteil 1277 • corps 1165 • embryologie 1157 • prématuré 1124 • embryon 1110 • ovule 1108 • prison 1096 • humaine 1084

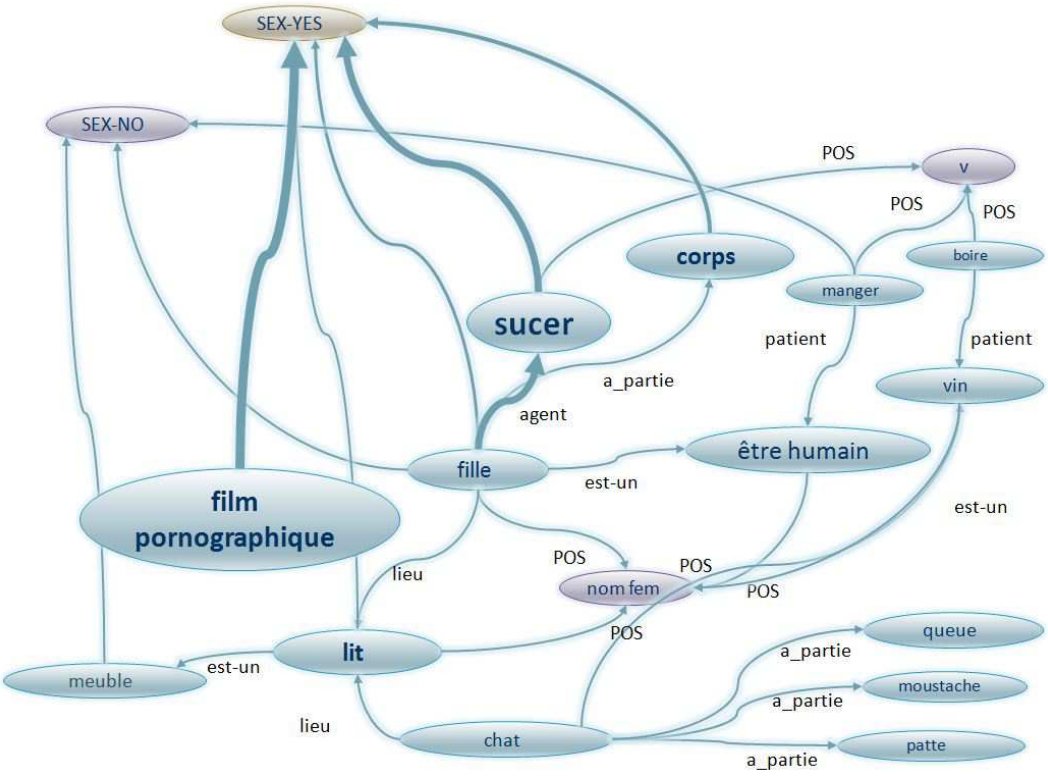


Figure 5: Effect of the propagation algorithm (terms related to sex are inflated).

• spermatozoïde 1039 • bouche 1011 · · ·

It reaches far beyond SexIt, as the whole network can be filtered out using any constraint/filter. Increasing the strength of the signal leads to better recall but lower precision. Considering that the JeuxDeMots network is a small world, with a diameter of around 6, the best fit (optimizing F1-score between recall and precision) is 3 (half the diameter). The higher the number of signals, the more confident we can be in the relative order of terms in the list, but the longer the computing time (for 100,000 signals of strength 3, the computing time on our servers is around 15 seconds).

When applied without any filter, with a starting term $x$, the result of the propagation algorithm converges toward the current state of the JeuxDeMots lexical network around $x$. The propagation algorithm can be alternatively used as a selector on constrained terms. For example, the following list has been obtained from the term *botanique* with the filter $\{x \xrightarrow{is-a} \textit{flower} \,\&\, x \xrightarrow{carac} \textit{red}\}$:

• rose 6349 • belle de nuit 4263 • digitale pourpre 2652 • orchis militaire 1972 • orchis singe 1904 • pigamon à feuilles d'ancolie 1778 • ancolie 1347 • bissap 1269 • œillet des rochers 1101 • lamier rouge 1096 • géranium brun 1065 • hedysarum des Alpes 1047 • cirse laineux 1047 • coquelicot 1041 • renouée persicaire 1038 • bugrane rampante 1038 • ciboulette 1029 • epilobe hérissée 1011 • myrtille 1011 • silène dioïque 1008 • orchis à feuilles larges 1005 • corydale à tubercule plein 1002 · · ·

If the needed data is not available, it has to be produced, for example using a game like SexIt, LikeIt, or ColorIt.

### 3.3. Evaluation

The French JeuxDeMots network contains, as of October 2013, around 300,000 terms, including 15,000 to 20,000 word usages and more than 6 million relations. The main game has been played more than 1.3 million times by more than 3,500 registered players.

For the SexIt data, more than 10,000 terms have been characterized as being related to sex (4,610) or not (8,158), for more than respectively 15,000 and 47,000 votes.

There is, to our knowledge, no equivalent resource for French (the free French WordNet WOLF (Sagot and Fišer, 2008) is no complete yet), and the Princeton WordNet (Fellbaum, 1998), for English, is quite different, with a smaller coverage (155,287 words as in its version 3). Above all, none of the WordNet-like resources allow for the dynamic update JeuxDeMots offers. Furthermore, links between terms in JeuxDeMots are weighted. The higher the weight, the stronger the activation of the relation between the terms.

Evaluating the JeuxDeMots resource is not easy, as there is no gold standard to compare it to. The same applies to SexIt. Indeed, apart from WOLF, there is no lexical resource available for French on which we can apply our approach. Moreover, even considering other languages, existing lexical networks do not provide the amount of information existing in JeuxDeMots, neither quantitatively nor qualitatively. Relevant information should include freely

associated ideas, common sense knowledge and conceptual representations. These information are included in JeuxDeMots with SexIt (among others) as a result of the players activity. However, the principle behind the presented algorithm can be applied to any lexical network.

It is however possible to manually evaluate the precision of the filtering service for some applications. There is no straightforward way to evaluate recall, again for lack of a gold standard. We therefore evaluated the performance of the filtering service on data filtered by the information collected from SexIt, as they are useful in applications like parental control, pornography contents filtering, etc.

In the lists given above as example and those in appendix, we ask people if the terms they contain are relevant. More precisely, for the list above, they have been asked to count the proper terms that are related to sex under the focus of biology, and those which are irrelevant. The same task have been applied on terms related to sex under the focus of pornography. In the previous list (which is only partially given here), 98% of the terms were relevant (*prison* is the sole exception). However, some terms are clearly polysemous, and might be relevant to the constraints for only one of their possible meanings.

One may ask, why not filtering terms from SexIt which have a majority of positive vote for sex (instead of having at least one vote)? The reason is that some terms, while being possibly related to sex, tend, in the eyes of people, not to be relevant in general, but could be, in a specific context (represented by the starting term).

## Conclusion

We presented here a semantic filtering service applied on a semantic network and freely available via the Web.[7] This service is based on dynamic language resources, obtained by crowdsourcing using GWAPs.

These results add to others in showing the fundamental shift in language resources building that is going on now (see (Chamberlain et al., 2013) for more details). If the use of crowdsourcing microworking platforms like Amazon Mechanical Turk can be questionable both in terms of ethics and of quality (Fort et al., 2011), GWAPs represent a change of paradigm that is only matched by the digitalization of resources.

However, creating an efficient GWAP involves developments that are out of reach for many. This is the reason why we intend to develop a platform allowing users to generate their own pseudo-game from JDM.

## 4. References


Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. In Gurevych, I. and Kim, J., editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

Cioffi, C., Pagliarecci, F., and Spalazzi, L. (2009). An anomaly-based system for parental control. In Smari, W. W. and McIntire, J. P., editors, *HPCS*, pages 193–199. IEEE.

Dong, Z. and Dong, Q. (2006). *HowNet and the Computation of Meaning*. WorldScientific, London.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2):413–420.

Habert, B., Nazarenko, A., Zweigenbaum, P., and Bouaud, J. (1998). Extending an existing specialized semantic lexicon. In *International Conference on Language Resources and Evaluation (LREC)*, pages 663–668, Granada, Spain.

Lafourcade, M. (2007). Making people play for lexical acquisition. In *7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand.

Lenat, D. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December.

Sagot, B. and Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco.

Siorpaes, K. and Hepp, M. (2008). Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60.

Thaler, S., Siorpaes, K., Simperl, E., and Hofer, C. (2011). A survey on games for knowledge acquisition. Technical report.

Varelas, G., Voutsakis, E., Petrakis, E. G. M., Milios, E. E., and Raftopoulou, P. (2005). Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proc. 7 th ACM Intern. Workshop on Web Information and Data Management (WIDM)*, pages 10–16. ACM Press.

von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM*, 51:58–67.

Vossen, P. (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.


## Appendix

Target term *pornographie* (pornography) with 10,000 signals sent:

• film X 5077 • sexe 4796 • film porno 4696 • film pornographique 4354 • pornographie 4005 • X 3169 • porno 3091 • érotisme 2642 • sexualité 2482 • film de boules 2445 • Rocco Siffredi 2444 • obscénité 2008 • cul 1874 • Katsuni 1754 • hardeuse 1728 • luxure 1714 • Clara Morgane 1686 • Hot d'or 1669 • obscène 1608 • film érotique 1598 • indécence 1580 • cinéma pornographique 1549 • lascivité 1452 • pornographique 1436 • sodomie 1397 • débauche 1341 • libertinage 1323 • pédo-pornographie 1296 • x 1267 • vulgaire 1265 • strip-tease 1262 • fellation 1242 • paillardise 1192 • film de cul 1173 • Nina Roberts 1173 • grossièreté 1139 • magazine de charme 1131 • classé X 1130 • Rocco 1126 • masturbation 1120 • indécent 1118 • im-

[7] http://www.jeuxdemots.org/propagate.php

pudique 1100 • hard>pornographie 1021 • travailleur du sexe 963 • éjaculation faciale 946 • queue 921 • Zara Whites 908 • acteur pornographique 887 • lasciveté 880 • faire bander 875 • cochonceté 869 • Traci Lords 829 • gonzo 820 • triple pénétration 807 • copulation 806 • actrice porno 788 • paraphilie 769 • sexe>organe sexuel 763 • excitation sexuelle 732 • bifle 732 • gorge profonde 725 • bouquin de cul 712 • partenaires multiples 663 • pénis 653 • baiser>faire l'amour 651 • Celia Blanco 629

Target term *reproduction* with 10,000 signals sent:

• reproduction 4015 • sexe 3360 • fécondation 2156 • accouplement 2123 • enfant 1966 • femelle 1751 • biologie 1630 • oeuf 1569 • parade amoureuse 1496 • homme 1476 • ovule 1433 • procréation 1431 • ovulifère 1385 • faire un bébé 1369 • copulation 1338 • femme 1316 • copuler 1311 • organe reproducteur 1292 • bébé 1264 • sperme 1239 • petit 1235 • vie 1211 • embryon 1208 • cellule sexuelle 1204 • testicule 1197 • double 1195 • gestation 1194 • caractères sexuels secondaires 1191 • enceinte 1184 • spermatozoïde 1173 • mâle 1131 • grossesse 1122 • lit 1114 • saison des amours 1108 • gamète 1094 • nature 1074 • saillie 1059 • sexualité 1033 • amphimixie 987 • méthode Ogino 981 • appareil reproducteur 967 • enceinte>grossesse 951 • reproduction>biologie 935 • cheval 926 • contraception 923 • couple 912 • ovulation 898 • accouplement>reproduction 885 • zoologie 860 • conception 842 • pénis 802 • gamète femelle 798 • nouveau-né 791

Target term *pratique sexuelle* (sexual practices) with 10,000 signals sent:

• pratique sexuelle 5928 • sexualité 4759 • sodomie 3550 • cunnilingus 3402 • fellation 3239 • masturbation 2739 • cravate de notaire 2241 • sexe 2136 • paraphilie 2112 • anal flower 1752 • sadisme 1681 • abstinence 1613 • bondage 1593 • masochisme 1519 • onanisme 1464 • pénétration anale 1396 • enculade 1235 • gorge profonde 1180 • 69 1107 • triple pénétration 1077 • voyeurisme 1057 • coït interrompu 1037 • saphophile 1030 • lesbophile 1022 • double pénétration 984 • sexe oral 983 • fist-fucking 937 • homosexualité 930 • langue de chat 930 • exhibitionnisme 921 • rapport sexuel 887 • triolisme 881 • nécrophilie 877 • coït 830 • autoérotisme 822 • missionnaire 822 • axilisme 817 • position sexuelle 815 • fessée érotique 806 • levrette 806 • coït anal 793 • plaisir 784 • sodomiser 772 • sado-maso 765 • sadomasochisme 763 • cunni 745 • abstinence sexuelle 743 • sexologie 742 • sexuel 711 • anulingus 708 • branlette espagnole 704 • coït vaginal 698 • éjaculation faciale 683 • sucer>fellation 678 • futution 677 • cuckolding 669 • ondinisme 668 • broute minou 654 • BDSM 647 • pipe>fellation 641 • préliminaire sexuel 631 • pénétration buccale 627 • bukkake 625 • fist-fucker 625 • pédophilie 622 • enculer 605 • faire jouir 601 • langue de chat>pratique sexuelle 596 • cuni 589 • sitophilie 589 • se faire péter la rondelle 589 • comportement sexuel humain 587 • sexuelle 580 • bifler 578 • donkey punch 571 • éjaculation 567 • feuille de rose 560 • orgasme 560 • érotisme 556 • pénis 553 • enculade>sodomie 542 • préliminaire 537 • sidérodromophilie 531 • irrumation 531 • autofellation 527 • sexe>rapport sexuel 523 • coprophile 522 • femme 519 • caresse anale 514 • domination 514 • chasteté 514 • fessée 508 • faire un cunnilingus 507 • gérontophilie 506 • cunnilinctus 506

Target term *maladie* (disease) with 10,000 signals sent:

• médicament 10900 • maladie 10166 • malade 9636 • fièvre 7956 • santé 7762 • médecine 7486 • douleur 6185 • virus 5904 • sida 4314 • pharmacie 3505 • SIDA 2858 • infection 2532 • chirurgie 2503 • génétique 2434 • mal 2335 • repos 1811 • maladie sexuellement transmissible 1781 • immunodéficience 1632 • préservatif 1595 • MST 1564 • corps 1550 • contracter 1512 • peur 1502 • syphilis 1443 • avoir mal 1378 • infirmière 1363 • hépatite 1133 • blennorragie 1103 • lit 1096 • maladie vénérienne 1040 • ampoule 1040 • longue 1008 • sexuellement transmissible 1001 • chaude-pisse 971 • sang 930 • sexe 894 • forme 879 • neurosyphilis 859 • hérédité 854 • gène 786 • congénital 770 • Durex 755 • traitement 721 • mal de Naples 710 • anatomie 707 • névrose 704 • addiction 691 • séropositif 688 • condom 683 • mauvaise humeur 664 • chromosome 644 • gonococcie 643 • dépression 635 • tête 618 • séropositivité 600 • biologie 568 • vérole 567 • vie 551 • ADN 535 • peau 534 • courbature 531 • vaginose 516 • HIV 511 • infirmier 506 • gonorrhée 495 • pilule 484 • VIH 475