Observatorio de la diversidad lingüística y cultural en la Internet
Observatoire de la diversité linguistique et culturelle dans l'Internet
Observatory of the Linguistic and Cultural Diversity in the Internet

https://OBDILCI.ORG/

Rapports sur le multilinguisme de la Toile No 3 :
Une caractérisation du Web francophone à partir d'une série de paramètres, en comparaison avec d'autres langues dominantes sur la Toile.

D. Pimienta, OBDILCI, 25/04/2025

1- CONTEXTE

Dataprovider.com gère une base de données rassemblant des informations relatives à un ensemble de sites web qui représentent près de la totalité du WWW (80 % si l'on compare aux chiffres fournis par Netcraft : 872 920 299 sur 1 101 431 853). Parmi les nombreuses informations conservées pour chaque site web, on trouve la langue principale unique du site (souvent extraite de la page d'accueil), paramètre que l'on peut désagréger avec une série d'autres paramètres qui permettent de caractériser la Toile francophone en comparaison avec d'autres langues dominantes sur la Toile.

Dataprovider.com a offert un accès de courtoisie à la base de données afin de permettre à l'OBDILCI de développer des statistiques utiles sur les langues en ligne, conformément à sa mission à but non lucratif. Ce rapport est le troisième d'une série de rapports sur les analyses réalisées par OBDILCI et obtenues grâce à la base de données de Dataprovider.com. La base de données de Dataprovider.com est impressionnante, à la fois par sa portée extraordinaire et peu commune, une grande partie de l'ensemble du WWW et par la facilité d'utilisation, la convivialité et la puissance de l'interface qui permettent des recherches en utilisant près de 100 paramètres différents avec des combinaisons illimitées et une réponse immédiate.

2- CHAMP D'ANALYSE DE L'ETUDE

Les paramètres concernés par cette étude sont les suivants :

- Le thème principal du site parmi une série de 27 possibilités
- Une métrique de l'impact économique du site, mesurée de 0 à 100
- Une métrique de la confiance accordée au site en termes de commerce électronique, mesurée de 0 à 100
- La taille du site en nombre de pages entre 1 et plus de 500
- Le nombre de liens rentrants vers le site, de 0 à plus de 1000

- La vocation d'affaire du site
- Le modèle d'affaire du site : b2b ou b2c.

Les données seront recherchées sur l'ensemble des sites francophones, sur les sites du domaine .fr, sur les sites francophones du domaine .fr et sur les sites logés en France, puis seront comparées avec les mêmes résultats dans d'autres langues, et, le cas échéant, aux sites de haut niveau correspondant aux pays associées aux langues (.it pour l'Italie, par exemple). Après avoir établi que les différences entre les différents abordages sont minimales, les résultats publiés sont ceux de « l'option langue du site est le français ».

3- METHODE

Le but de cette étude est de tenter une **caractérisation de la Toile francophone** en comparaison avec d'autres langues très présentes sur la Toile.

Chacun des paramètres recevra un traitement particulier qui permettra cette caractérisation. De l'ensemble des résultats pour la Toile francophone, et des comparaisons avec les mêmes résultats pour d'autres langues, devraient émerger des enseignements sur les forces et les faiblesses du web francophone et en synthèse une caractérisation de la Toile francophone.

Le paramètre *thématique* (« web topics ») déterminera une sorte de **signature thématique** en mesurant, pour chacun des thèmes, les *écarts de pourcentages* entre la Toile francophone et l'ensemble de la Toile ou avec une autre langue.

La liste complète des thèmes est présentée dans le tableau suivant, la première colonne, en anglais, dans sa version originale, la deuxième colonne la traduction libre tenant en compte du contexte.

Tableau 1 : Liste des thèmes possibles associés avec les sites web

WEB TOPICS	**THEMES**
Adult	Pornographie
Art & Design	Art et design
Beauty & Personal care	Soins de beauté et personnels
Business & Careers	Affaires et carrières
Construction, Repair & Installation	Construction, réparation etc.
Education	Education
Electronics & Hardware	Matériels électroniques
Entertainment	Divertissement
Fashion	Mode
Financial	Finance
Food & Drinks	Gastronomie
Government & Society	Gouvernement & Société
Health & Medical	Santé et médecine
Home & Garden	Maisons et jardins
Industrial	Industrie
Legal	Juridique
Marketing & Advertisement	Marketing et publicité
Media	Médias

Nature	Nature
Real estate	Immobilier
Religious	Religions
Science	Science
Software & Data	Logiciels et données
Sports	Sports
Telecom	Telecom
Transport	Transport
Travel & Tourism	Voyage et tourisme

L'assignation thématique est déterminée automatiquement par un algorithme qui semble fonctionner dans la grande majorité des langues reconnues par la base de donnéees.

Les travaux montrent que chaque langue a une signature particulière très différente de celle des autres langues, signature cohérente, dans le cas du français que l'on explore l'ensemble des sites francophones, l'ensemble des sites du domaine Internet national de la France .fr ou les sites francophones du domaine .fr, ou enfin pour les sites situés en France. Certains résultats, comme l'importance de la *gastronomie* dans la signature francophone, seront attendus mais d'autres pourront réserver des surprises, comme la proportion relativement faible de sites sur *l'éducation* ou la *finance*.

Il est très important pour analyser les résultats de comprendre qu'il s'agit essentiellement de donnéees **quantitatives** (le nombre de sites web classifiés sous telle ou telle thématique) et statiques (résultats indépendants du trafic vers les sites comptabilisés), mais en aucun cas qualitative ou dynamique. Les donnéees ne renseignent pas sur la qualité, la découvrabilité ou la puissance de trafic des sites comptabilisés : tous les sites sont comptabilisés comme une unité, indépendamment de leur qualité, place dans les résultats de recherches ou nombre de visites.

Il est également important de prendre en compte les possibilités de **biais** incontrôlables des résultats établis (voir plus loin le chapitre biais).

Dans un premier temps, il est établi, comme référence de calcul, la signature thématique générale de l'ensemble de la Toile, en termes de pourcentage de sites dans chaque catégorie thématique. La base de donnéees comprend un peu plus de 885 millions de sites mais le travail est réalisé sur le sous-ensemble des sites dit « développés », c'est à dire des sites qui existent vraiment[1] et sont disponibles (excluant les sites en attente de développement et les sites en réserve – en anglais *parked domains*).

Le tableau ci-dessous se lit ainsi : la base de données contient 83 millions de sites développés ; pour un sous ensemble de 55 millions d'entre eux (66%) le paramètre thématique a été établi. Certains sites se voient attribués plus d'un thème, pour cette raison le total des pourcentages dépasse 100% (il y a sur l'ensemble 24,5% de mentions thématiques non uniques). Le thème le plus présent sur la Toile est celui du *divertissement*, avec 7,8 millions de sites, soit 14%

[1] C'est un phénomène chronique, que l'on peut vérifier sur les donnéees fournies par Netcraft (source : https://www.netcraft.com/blog/january-2025-web-server-survey/) qu'il existe plus de 5 fois moins de sites en fonctionnement (dit « actifs ») que de sites attribués (en janvier 2025, 195 millions versus 1161 millions). Cet écart ne semble pas diminué au cours du temps.

d'entre eux. Les moins présents sont les sites pornographiques et scientifiques[2] (respectivement 1,04% et 0,99%).

Tableau 2 : Répartition en pourcentage des thèmes pour l'ensemble de la Toile

TOTAL PARAMETRES	**55 045 566**	**124,53%**
TOUTE LA TOILE	**83 543 409[3]**	**66%**
THEMES	**NOMBRE DE SITES**	**%**
Divertissement	7 875 236	14,31%
Construction, réparation, etc.	6 011 329	10,92%
Logiciels et données	5 378 362	9,77%
Art et design	4 571 331	8,30%
Soins de beauté et personnels	3 890 032	7,07%
Santé et médecine	3 877 682	7,04%
Gastronomie	3 580 745	6,51%
Maisons et jardins	3 215 421	5,84%
Marketing et publicité	2 432 946	4,42%
Mode	2 428 324	4,41%
Médias	2 396 200	4,35%
Voyage et tourisme	2 282 722	4,15%
Education	2 223 561	4,04%
Industrie	2 062 449	3,75%
Transport	2 046 364	3,72%
Finance	2 013 320	3,66%
Sports	1 943 768	3,53%
Immobilier	1 893 414	3,44%
Matériels électroniques	1 518 179	2,76%
Gouvernement & Société	1 508 801	2,74%
Religions	1 204 763	2,19%
Juridique	1 090 409	1,98%
Nature	675 064	1,23%
Affaires et carrières	669 948	1,22%
Telecom	641 264	1,16%
Pornographie	574 655	1,04%
Science	542 299	0,99%

Dans un second temps, le même tableau est établi, après avoir fixé le paramètre langue = français.

[2] Triste ironie de l'évolution pour un phénomène qui est né et a mûri dans le monde de la recherche avant de se généraliser vers le grand public.

[3] La base de donnéees de DataProvider est mise à jour chaque semaine alors que cette étude a duré plusisurs semaines. Cela explique des différences marginales entre certaines valeurs. Pour l'ensemble des travaux ces différences n'affectent pas les valeurs des pourcentages pour les deux chiffres après la virgule et ne sont donc pas à prendre en compte.

Tableau 3 : Répartation en pourcentage des thèmes pour la Toile francophone

LANGUE=FRANÇAIS TOTAL SITES>	**3 292 841**	**71%**
SITES PARAMETRISES>	**2 335 218**	**121,54%**
THEMES	**NOMBRE**	**%**
Pornographie	10 543	0,45%
Art et design	214 601	9,19%
Soins de beauté et personnels	182 272	7,81%
Affaires et carrières	28 552	1,22%
Construction, réparation etc.	301 368	12,91%
Education	67 394	2,89%
Matériels électroniques	57 523	2,46%
Divertissement	275 381	11,79%
Mode	104 898	4,49%
Finance	40 721	1,74%
Gastronomie	237 506	10,17%
Gouvernement & Société	89 934	3,85%
Santé et médecine	150 899	6,46%
Maisons et jardins	161 887	6,93%
Industrie	41 426	1,77%
Juridique	37 487	1,61%
Marketing et publicité	74 845	3,21%
Médias	66 166	2,83%
Nature	30 035	1,29%
Immobilier	49 232	2,11%
Religions	28 446	1,22%
Science	15 672	0,67%
Logiciels et données	136 453	5,84%
Sports	138 207	5,92%
Telecom	10 586	0,45%
Transport	100 697	4,31%
Voyage et tourisme	185 504	7,94%

Ensuite, le rapport entre le pourcentage des sites francophones est divisé par la même valeur pour la Toile entière et une légère correction est établie pour tenir compte des pourcentages différents de thèmes non uniques (multiplication par 124,53/121,54) est ainsi nous obtenons la signature francophone en termes d'écart par rapport à l'ensemble :

Tableau 4 : Comparaison de la Toile francophone avec l'ensemble de la Toile

THEMES	RATIO FR VS TOUS	RATIO TOUS VS FR
Voyage et tourisme	1,96	0,51
Sports	1,72	0,58
Gastronomie	1,60	0,62
Gouvernement & Société	1,44	0,69
Maisons et jardins	1,22	0,82
Construction, réparation etc.	1,21	0,83
Transport	1,19	0,84
Art et design	1,13	0,88
Soins de beauté et personnels	1,13	0,88
Nature	1,07	0,93
Mode	1,04	0,96
Affaires et carrières	1,03	0,97
Santé et médecine	0,94	1,06
Matériels électroniques	0,92	1,09
Divertissement	0,84	1,18
Juridique	0,83	1,20
Marketing et publicité	0,74	1,35
Education	0,73	1,37
Science	0,70	1,43
Médias	0,67	1,50
Immobilier	0,63	1,59
Logiciels et données	0,61	1,63
Religions	0,57	1,75
Finance	0,49	2,05
Industrie	0,49	2,06
Pornographie	0,44	2,26
Telecom	0,40	2,51

Ce tableau, qui représente la signature thématique du Web francophone, se lit ainsi : il y a presque deux fois plus de sites en français sur les *voyages et le tourisme* que la proportion mondiale ; il y a deux fois moins de sites en français sur *la finance ou l'industrie* que la proportion mondiale. Les thèmes qui mobilisent, en proportion, le plus de sites en français sont dans l'ordre : *tourisme, sports, gastronomie et gouvernement et société*. Les thèmes pour lesquels la proportion de sites en français est très inférieure à la proportion mondiale sont dans l'ordre : *télécommunications, pornographie, industrie, finance et religion.*

Mais n'oublions pas que le critère est purement **quantitatif**. Par exemple, il est probable que le score de Telecom soit corrélé, entre autres, au nombre de fournisseurs de service lequel est très variable d'un pays à l'autre.

La même méthode a été utilisée avec le domaine .fr et ensuite avec les sites francophones à l'intérieur du domaine fr et finalement à l'intérieur des sites logés en France : les résultats sont extrêmement proches dans tous les cas de figure.

Pour compléter la comparaison et avoir une meilleure idée de la signature thématique francophone dans la Toile, la méthode consiste à faire la même évaluation pour une série de langues qui sont prédominantes sur la Toile et également de faire, pour chacune de ses langues une comparaison avec le français.

Les langues suivantes ont été traitées : *anglais, espagnol, hindi, portugais, chinois, russe, japonais, allemand, hollandais et italien.*

Enfin, de manière à relativiser les résultats, il a été établi, pour chaque thème, le classement des langues en termes de pourcentages, du plus haut vers le plus bas. Ainsi, pour les sites francophones, le thème le plus traité en proportion au reste des sites, est *tourisme et voyage.* Il est intéressant de noter le classement pour *tourisme et voyage* montre le français en deuxième position dans le classement, derrière l'italien.

La comparaison entre les résultats de pourcentages de thèmes par langue et les résultats de pourcentage de langues par thème a provoqué de nombreuses interrogations sur la validité des résultats en raison d'inconsistances ou de contradictions (par exemple, pour le web francophone, le thème *télécom* est un des plus faibles par rapport à la moyenne dans un cas et pourtant il place le français en tête des langues dans l'autre cas). Le rôle important joué par l'anglais sur les moyennes pondérées fournit un début d'explication mais c'est insuffisant pour lever les doutes.

Les calculs prenaient en compte les pourcentages différents de cas où le même site web se voit attribué plus d'un thème et établissaient une correction pour rendre justice aux différences. Par contre il n'avait pas été pris en compte le pourcentage, par langue, de sites étant traités pour le paramètre thématique. L'étude de ces pourcentages a fourni l'explication et a conduit **à revoir complètement la méthode.**

En effet, seules 19 langues ont des pourcentages de sites traités supérieurs à 65%. Le reste des langues reçoit un traitement dans une proportion très restreinte, inférieure à 10%, voire 1%, dans la majorité des cas, y compris pour des langues importantes comme l'arabe, l'indonésien, l'ukrainien, le tchèque ou le vietnamien. L'ensemble des sites traités pour les langues autres que les 19 mentionnées représente moins de 1% des sites traités. Dans ces conditions, le seul traitement comparatif valable concerne uniquement ces 19 langues. Les deux paramètres (% de sites traités par langue et % de thèmes par langue) doivent, de plus, être pris en compte pour établir les corrections de manière à remettre à égalité toutes les langues[4].

[4] Les corrections se feront simplement par règle de trois, en multipliant les données par les pourcentages moyens divisés par le pourcentage propre à la langue.

Tableau 5 : Couverture thématique par langue pour les 19 langues les mieux traitées

LANGUE	NB TOTAL SITES	NB SITES TRAITES	% TRAITES	% THEMES
Allemand	5 913 273	3 862 498	65,3%	119,38%
Anglais	41 112 849	30 982 345	75,4%	124,28%
Chinois	3 970 596	2 719 692	68,5%	115,62%
Coréen	522 903	343 532	65,7%	119,88%
Danois	341 011	267 459	78,4%	129,32%
Espagnol	3 679 518	2 575 560	70,0%	119,95%
Finlandais	229 299	175 173	76,4%	128,21%
Français	3 292 841	2 271 577	69,0%	121,54%
Hébreu	176 679	132 339	74,9%	129,18%
Hindi	77 957	53 725	68,9%	114,63%
Hollandais	1 743 766	1 337 464	76,7%	124,81%
Italien	1 678 656	1 218 505	72,6%	123,52%
Japonais	3 009 688	2 102 090	69,8%	120,07%
Norvégien	208 576	155 892	74,7%	127,87%
Polonais	866 016	658 047	76,0%	124,96%
Portugais	2 427 389	1 758 061	72,4%	124,46%
Russe	2 306 354	1 834 718	79,6%	134,50%
Suédois	515 669	394 079	76,4%	129,33%
Turc	962 390	733 605	76,2%	127,72%
TOTAL/MOY	**73 035 430**	**53 576 361**	**73,5%**	**124,53%**

Les comparaisons seront donc établies entre les 19 langues mentionnées, en utilisant le classement pour chaque thème. L'indicateur choisi pour les comparaisons est ainsi le classement de 1 à 19 de chaque langue pour chaque thème. Cet indicateur permettra d'obtenir des résultats fiables et exploitables.

Pour le paramètre *impact économique* (economic footprint) la base de données fournit le nombre de sites par tranche : 0 - <10, 10 - <20, 20 - <30, 30 - <40, 40 - <50, 50 - <60, 60 - <70, 70 - <80, 80 - <90, 90 - <101. Pour chaque groupe de sites mesurés, nous établissons la moyenne pondérée en multipliant le nombre de sites par tranche, respectivement par les valeurs 5, 15, 25, 35, 45, 55, 65, 75, 85, 95 et en divisant la somme de ces multiplications par le nombre total de sites. Cette valeur pondérée représente le poids global pour le groupe traité et permet ensuite les comparaisons avec d'autres groupes. Le calcul est réalisé par langue er par domaine national du pays et la moyenne est retenue. Pour certaines langues, la valeur de deux domaines est retenue (par exemple, pour le portugais, Brésil et Portugal).

Pour le paramètre *degré de confiance pour commerce électronique* (« E.com Trust Grade »), la base de données répond avec une échelle de 5 entre A (le plus fort) et F (le plus faible). Une opération de pondération est réalisée en transformant les lettres par des valeurs 100, 80, 60, 40, 20, 0. La valeur pondérée obtenue permet de comparer les groupes de sites web.

Pour le paramètre *taille* (« Pages ») la réponse de la base de donnéees est le nombre de pages sites dans les segments suivants : 1 - <2, 2 - <10, 10 - <25, 25 - <50, 50 - <100, 100 - <200, 200 - <300, 300 - <400, 400 - <500. La moyenne pondérée est faite avec les valeurs : 1, 5, 12,5,

37,5, 75, 150, 250, 350, 450. Un développement original de ce paramètre en le combinant avec les donnéees démolinguistiques et de connectivité à l'Internet qui sont présentes dans le modèle d'OBDILCI a permis de calculer les valeurs d'indicateurs clefs de ce modèle et d'établir la comparaison entre les deux modèles, OBDILCI et DataProvider, de manière à mesurer les biais linguistiques de la base de données.

Le paramètre *âge du site* (« website age ») a été abandonné après la découverte de problèmes qui ont été signalés à l'entreprise.

Pour le paramètre *liens entrants* (incoming links) la base de données donne le total de sites par tranche : 0 - <1, 1 - <2, 2 - <5, 5 - <10, 10 - <100, 100 - <1K, > 1K et nous réalisons une pondération avec les valeurs 0, 1, 2,5, 7,5, 45, 450, 1500 pour établir le nombre moyen de liens par groupe traité, à des fins de comparaisons.

Pour le paramètre de *taux d'actualisation* (changes) la base de données travaille avec une échelle de 7 niveaux, entre nul et extrême, et nous réalisons la pondération avec les pourcentages respectifs 0%, 10%, 20%, 45%, 60%, 75% et 95%. Le constat de différences significatives entre les résultats par langue et les résultats par domaine de haut nveau des pays correspondant à ces langues ont conduit finalement à disqualifier ce paramètre pour cette étude.

Pour le paramètre *affaire* les réponses obtenues sont d'abord les pourcentages de sites classifiés « affaires » et ensuite les pourcentages respectifs de sites b2b et b2c. De simples moyennes de ces pourcentages par langues seront ainsi comparés.

4- BIAIS

Les résultats sont susceptibles de présenter plusisurs biais, dont la plupart ne sont pas maitrisable par l'auteur de ce rapport et/ou sont indéterminés, en l'absence de description détaillée des algorithmes sous-jacents.

4.1 Biais de détection de langue.

L'analyse des résultats de la base de données dans les deux études précédentes[5] a mis en évidence que, pour certaines langues, l'algorithme de détection des langues utilisé par dataProvider.com peut être défaillant dans des proportions notables. Cependant, ce n'est pas le cas pour les langues de travail de cette étude, à l'exception peut-être de l'hindi qui est faiblement représenté dans la base par rapport à la réalité de sa présence (mais ce n'est pas forcément un problème de détection).

4.2 Biais potentiels de l'algorithme d'assignation de thématiques.

La liste des thèmes est fixée par DataProvider et nous devons la prendre telle quelle (nous aurions préféré par exemple avoir un thème général « culture »). Nous ne connaissons pas les détails de l'algorithme qui détermine le ou les thèmes à assigner à un site web. Nous avons fait un contrôle statistique en vérifiant sur un large échantillon de sites pris au hasard si l'assignation est correcte et les résultats sont plutôt positifs, avec un taux d'erreur détecté inférieur à 10%.

[5] Voir les études (en anglais) à https://www.obdilci.org/projets/autres/mlreports-2/

Nous ne connaissons pas le critère qui permet d'assigner plus d'un thème au même site. Nos tests aléatoires ont montré que, par exemple, sous la rubrique « gastronomie » on peut trouver aussi bien des restaurants que des sites de ventes de produits alimentaires ou des sites de recettes. De même, sous la rubrique « mode » on trouvera de simples boutiques de vêtements aussi bien que des sites plus directement concernés par la mode. Un site qui vend de la lingerie féminine ou des maillots de bain pour femme et qui comporte des photos en maillots de bain pourra être classé comme mode et pornographie (traduction un peu exagérée de « adult ») en même temps. Sous la rubrique *gouvernement et société* apparaissent des sites gouvernementaux, des associations et des institutions à fonction sociétale. Une association du type « restaus du cœur » pourrait apparaitre sous la double assignation *gouvernement et société* et *gastronomie*.

Il est important de comprendre cette logique pour interpréter les résultats et de savoir que le résultat des tests aléatoires des assignations (une centaine de site examinés pour chaque critère) apparait tout à fait fiable et cohérent. Cependant, il faut le répéter, il n'y a aucun jugement qualitatif, la seule donnée fiable est la **quantité** de sites assignés à chaque thématique et un site essentiel compte autant qu'un site insignifiant, et il en est de même pour un site très visité par rapport à un site rarement visité.

Pour chacun des autres paramètres, nous ne connaissons pas les détails de l'algorithme de détermination des classements de sites, mais là encore, des tests aléatoires ont montré un faible taux d'erreur.

5- RESULTATS

5.1 Thématiques

En synthèse, l'histogramme ci-dessous présente les résultats du classement du web francophone en comparaison avec les 18 autres langues. En dégradé, du rouge vers le bleu, les thèmes pour lesquels le web francophone a un classement parmi les derniers jusqu'aux classements parmi les premiers ; au milieu en gris, les thèmes proches de la moyenne. Un classement parmi les premiers indique que le pourcentage de sites pour ce thème est supérieur aux pourcentages de la plupart des langues.

Les thèmes les plus « forts » pour la Toile francophone (c'est à adire avec des classements parmi les premiers) sont *gastronomie, voyage et tourisme*, *nature*, *sports* et *art et design*. Les thèmes les plus « faibles » sont *finance, logiciels et données*, *éducation* et *religions*.

Le tableau en séquence indique également les quatre langues qui sont en tête de classement, pour chaque thème. En annexe, les résultats complets par thème sont consignés. La consultation des tableaux en annexe permet de voir des écarts parfois très importants entre la première langue et les suivantes. C'est le cas, par exemple, pour *industrie* qui montre en tête sites le chinois avec plus de 22% des sites, contre un peu moins de 9% pour le turc ; ou bien *télécom*, qui montre 1,7% des sites anglophone et en séquence les pourcentages décroissent rapidement à partir de 0,8% ; ou bien encore *mode* qui montre 22% des sites pour le coréen pour descendre à 4% pour la langue suivante, le portugais.

Ce qui est remarquable, c'est la totale diversité des résultats : chaque langue possède une signature thématique qui lui est propre et très différente du reste des langues.

Tableau 6 : Signature thématique du web francophone en termes de classement

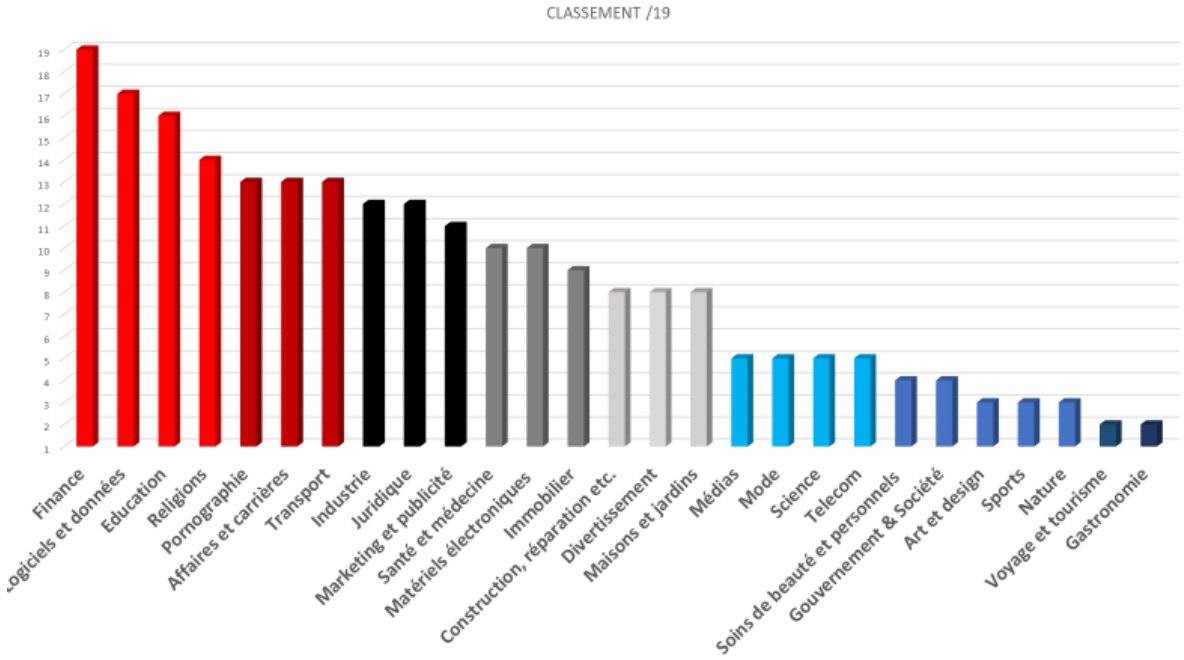


Tableau 7 : Rang associé à chaque thème dans le web francophone

THEMES	RANG /19	CINQ PREMIERES LANGUES
Finance	19	Portugais, norvégien, polonais, hébreu, finlandais
Logiciels et données	17	Russe, chinois, norvégien, suédois, espagnol
Education	16	Polonais, japonais, portugais, hébreu, hindi
Religions	14	Hindi, hébreu, coréen, anglais, allemand
Pornographie	13	Chinois, turc, japonais, hindi, russe
Affaires et carrières	13	Norvégien, suédois, hindi, portugais, danois
Transport	13	Turc, russe, finlandais, polonais, portugais
Industrie	12	Chinois, turc, espagnol, japonais, polonais
Juridique	12	Hébreu, portugais, polonais, allemand, italien
Marketing et publicité	11	Portugais, hébreu, anglais, allemand, chinois
Santé et médecine	10	Allemand, japonais, danois, hébreu, espagnol
Matériels électroniques	10	Chinois, portugais, norvégien, danois, coréen
Immobilier	9	Anglais, hébreu, portugais, espagnol, italien
Construction, réparation etc.	8	Turc, finlandais, suédois, russe, polonais
Divertissement	8	Hindi, chinois, japonais, allemand, portugais
Maisons et jardins	8	Coréen, hébreu, danois, hollandais, polonais
Médias	5	Anglais, hébreu, allemand, italien, **français**
Mode	5	Coréen, portugais, espagnol, anglais, **français**
Science	5	Russe, anglais, portugais, italien, **français**
Telecom	5	Anglais, portugais, hébreu, russe, **français**
Soins de beauté et personnels	4	Japonais, danois, finlandais, **français**, suédois

Gouvernement & Société	4	Hindi, russe, hollandais, **français**, italien
Art et design	3	Allemand, italien, **français**, anglais espagnol
Sports	3	Allemand, finlandais, **français**, danois, hollandais
Nature	3	Anglais, norvégien, **français**, espagnol, japonais
Voyage et tourisme	2	Italien, **français** norvégien, allemand, coréen
Gastronomie	2	Italien, **français**, coréen, allemand, japonais

Analyse des résultats pour le français

Que certains thèmes, comme *gastronomie* ou *art et design*, soient classés en tête du classement ne représente pas de surprise en soi mais certains thèmes placés en fin de classement interrogent. Il faut noter en passant dans le cas de *gastronomie* l'écart important des deux premières langues, italien et français, très proches l'une de l'autre, sur les suivantes. Il n'y a pas malheureusement de thème « *culture* » et celui qui se rapproche le plus est *art et design*[6] *;* en l'absence d'autres thèmes liés à la culture (musique, poésie, théatre, cinéma…) il est raisonnable de penser qu'il reflète au mieux ce secteur mais un certain doute peut subsister sur sa capacité à englober tous les aspects culturels.

Il est, par contre, plus intéressant d'analyser les thèmes qui apparaissent en bas du classement et là les surprises sont possibles, en particulier, *éducation*, en classement 16/19 interroge. Est-ce que cette place en fin de classement est un signal de phénomènes qui mériteraient analyse ? Exemple de questions que l'on peut se poser et qui dépasse le cadre de cette étude :

- Est-ce que ce résultat s'explique parce que l'éducation en France est essentiellement un secteur de l'état ? La place du privé étant restreinte par rapport à d'autres pays pourrait expliquer un écosystème moins dynamique et décentralisé.
- Les institutions de l'éducation supérieure françaises ont toutes un site web associé, ce n'est pas le cas de l'éducation primaire et secondaire ; est-ce différent dans d'autres pays ?

On peut se poser les mêmes questions pour le secteur de la *santé* pour lequel la francophonie se place seulement dans la moyenne.

Le classement très moyen du web en français pour le thème *industries* le place toutefois à côté de l'anglais, ce qui est rassurant, mais très loin des deux premiers, chinois et turc, qui ont une avance considérable (voir tables en annexe).

La toile francophone prend la dernière place de ce classement dans le secteur de la *finance*. Est-ce le reflet d'un écosystème autour de la finance qui n'a pas la richesse et la diversité que l'on connait dans d'autres pays ? Pourquoi le portugais est-il en première place ? La réponse est probablement à chercher au Brésil.

Il n'est par contre pas surprenant de constater que le nombre de sites consacré aux *religions* est proportionnellement plus faible dans la Toile francophone comparée à celle d'autres langues comme l'hindi et l'hébreu.

[6] Nous ne considérons pas que le thème divertissement, un peu fourre-tout, est de loin le thème qui reçoit pour l'ensemble de la Toile le plus haut pourcentage de sites (plus de 14% des sites de la Toile sont dans cette catégorie) puisse être représentatif de la culture.

Analyse des résultats pour les autres langues

Il est intéressant d'observer les situations où une langue se distingue particulièrement.

Pour la *mode*, le coréen montre une avance exceptionnelle sur les langues qui la suivent, avec 22% de sites sur ce thème alors que le portugais, en seconde position montre un score de 6%. Il semble effectivement que la Corée du sud se distingue dans ce domaine, tant culturellement, par une attitude générale de grande sensibilité à ce thème, que par un écosystème économique très large et distribué[7]. Pour le portugais, c'est probablement l'importance de la mode au Brésil qui lui vaut la seconde place.

Le thème *divertissement* est, en moyenne mondial, celui du plus grand pourcentage de sites. Les deux premiers ont des scores gigantesques : l'hindi avec 78% des sites et le chinois avec 35%, suivis par le japonais avec 21% et l'allemand avec 14%. Les autres langues vont de 12 à 6% avec le français dans la partie haute et le polonais et l'hébreu en fin de classement.

Les résultats pour les sites de *pornographie* sont assez comparables, avec le chinois détaché devant avec 8%, les 4 suivants entre 2,5% et 1%. Les langues européennes loin derrière, entre 0,8 et 0,4%, le coréen en fin de classement avec 0,3%

Coefficient de variation

Le coefficient de variation permet de visualiser les thèmes pour lesquels les résultats en pourcentage présentent les écarts les plus importantes. Il doit être pris en compte au moment de l'analyse, thème par thème. La valeur haute par exemple pour les sites de pornographie montre qu'il y a des écarts maximums entre les langues, certaines langues comme le chinois, le turc, le japonais ou le russe avec des pourcentages très hauts, d'autres langues, comme les langues européennes, avec des pourcentages très bas, ce qui se reflète dans une variance forte. En revanche, les écarts de pourcentage entre les pourcentages de sites du thème éducation sont extrêmement faibles et donc les différences dans le classement peuvent être marginales.

Tableau 8 : Coefficient de variation associé à chaque thème

	CV
Pornographie	**1,53**
Telecom	**1,17**
Industrie	**1,13**
Mode	**1,08**
Divertissement	**1,06**
Juridique	**0,79**
Science	**0,69**
Médias	**0,68**
Affaires et carrières	**0,63**
Religions	**0,61**
Art et design	**0,59**
Sports	**0,54**

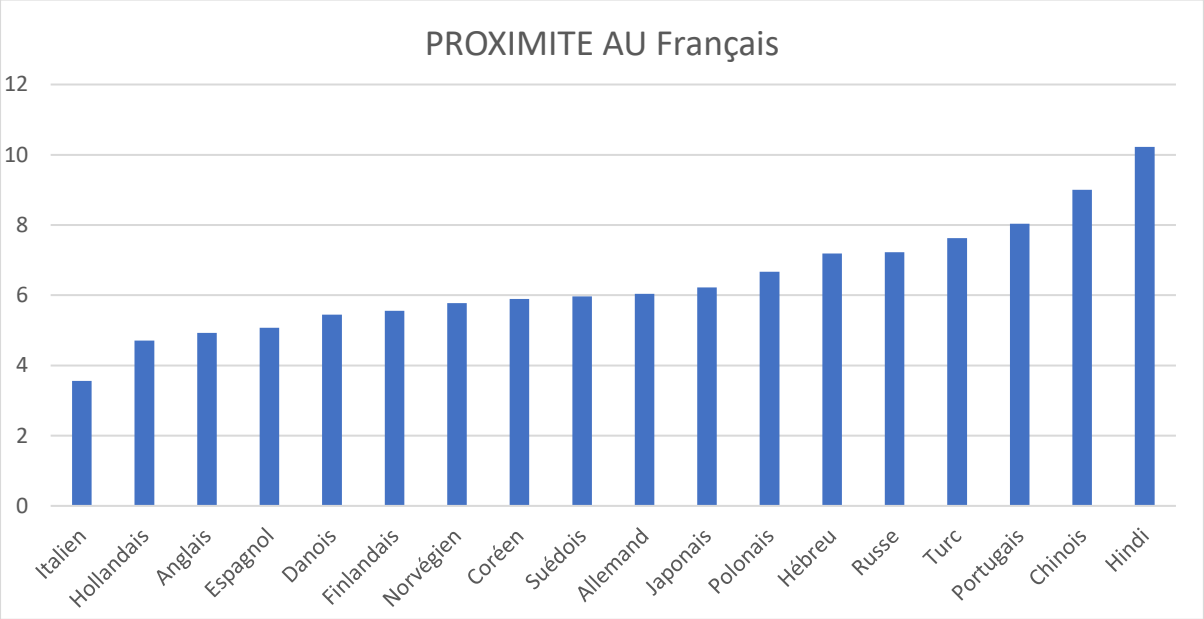
[7] Quelques références directes ou indirectes sur la prévalence de la mode en Corée du sud : https://unctad.org/fr/news/le-modele-k-pop-sinspirer-des-industries-creatives-sud-coreenne
https://fr.fashionnetwork.com/news/Coree-du-sud-terre-d-opportunites-pour-la-mode-francaise,1359219.html
https://fashionunited.fr/actualite/mode/la-coree-du-sud-est-elle-l-avenir-du-luxe/2023050532050

Maisons et jardins	**0,52**
Nature	**0,52**
Immobilier	**0,51**
Voyage et tourisme	**0,49**
Transport	**0,48**
Construction, réparation etc.	**0,45**
Soins de beauté et	**0,45**
Gouvernement & Société	**0,45**
Finance	**0,44**
Matériels électroniques	**0,43**
Marketing et publicité	**0,42**
Gastronomie	**0,40**
Santé et médecine	**0,40**
Logiciels et données	**0,39**
Education	**0,37**

Mesure de la distance au web francophone

Nous avons constitué la matrice des classements avec toutes les langues et tous les thèmes ce qui permet des traitements comparatifs complets sur les données. Ainsi, à partir de cette matrice nous avons établi la « distance » au français en mesurant la moyenne des écarts absolus entre les autres langues et le français. Sans surprise, le web italien apparait comme le plus proche du web francophone et les web chinois et hindi apparaissent les plus éloignés.

Tableau 9 : Classement des langues par la proximité au web en français en termes de signature thématique



Après la Toile italienne, les Toiles hollandaises, anglaises et espagnoles sont les moins éloignées de la Toile française, en ce qui concerne les thématiques. Un éloignement moyen de 3,56 positions, entre les classements français et italien, n'empêche bien entendu pas certaines divergences, comme le montre l'extrait suivant du tableau de comparaison entre français et italien, lequel ne conserve que les écarts supérieurs à 5 positions :

Tableau 10 : Ecarts les plus hauts entre toile française et italienne

THEME	FR	IT	FR-IT
Beauté et soins personnels	4	15	-11
Nature	3	13	-10
Religions	14	6	8
Juridique	12	5	7
Affaires et carrières	13	7	6

La conclusion principale de cette étude est l'existence de signatures thématiques clairement marquées et distinctes pour chaque espace linguistique, répondant en toute probabilité, aux critères socio-culturels et économique associés, dans le monde numérique, à chaque langue. En annexe, les signatures thématiques de l'anglais, l'allemand, l'italien, le japonais, le portugais, l'espagnol et le russe sont présentées.

L'analyse de la relation de proximité entre toutes les langues confirme cette conclusion. Le tableau suivant reprend la matrice des distances entre les langues (voir en annexe) et pour la rendre plus lisible, remplace les valeurs originales par une échelle proportionnelle de 0 à 100. L'interprétation se fait à partir du barème suivant :

0-15 : Très proche - *15-30 : Assez proche* - 30-50 : Plutôt proche
50-70 : Plutôt éloigné - **70-85 : Assez éloigné** - 85-100 : Très éloigné

Tableau 11 : Distance normée des signatures thématiques entre chaque langue

	CH	DA	HL	AN	FI	FR	AL	HE	HI	IT	JP	CO	NO	PO	PT	RU	ES	SU	TK
Chinois	X	x	x	x	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x
Danois	59	x	x	x	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x
Hollandais	**77**	9	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Anglais	**74**	57	43	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Finlandais	**74**	*19*	*18*	53	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Français	**82**	*28*	*17*	*21*	*30*	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Allemand	80	42	*24*	38	*19*	37	x	x	x	x	x	x	x	x	x	x	x	x	x
Hébreu	89	46	42	19	63	54	43	x	x	x	x	x	x	x	x	x	x	x	x
Hindi	*29*	69	67	91	**72**	100	**79**	93	x	x	x	x	x	x	x	x	x	x	x
Italien	**79**	49	*25*	29	47	0	*25*	47	**86**	x	x	x	x	x	x	x	x	x	x
Japonais	57	42	51	61	34	40	36	59	60	59	x	x	x	x	x	x	x	x	x
Coréen	58	39	40	38	58	35	49	39	59	*16*	63	x	x	x	x	x	x	x	x
Norvégien	59	*28*	*29*	43	*22*	33	45	57	64	41	50	35	x	x	x	x	x	x	x
Polonais	87	45	48	57	*28*	47	53	33	89	36	36	46	36	x	x	x	x	x	x
Portugais	63	58	69	31	65	67	56	35	**78**	53	58	44	58	35	x	x	x	x	x
Russe	37	66	41	62	47	55	**70**	56	61	52	47	63	49	34	54	x	x	x	x
Espagnol	**73**	37	23	8	39	*23*	*28*	23	98	13	44	*27*	45	*22*	*18*	44	x	x	x
Suédois	**70**	8	13	51	**6**	36	37	57	63	42	38	47	15	38	62	38	32	x	x
Turc	38	43	57	65	42	61	**74**	41	60	56	34	41	48	*17*	43	11	38	39	x

La relation de proximité n'est pas transitive. Cela signifie que si L1 et L2 sont assez proches de L3 cela n'implique pas nécessairement que L1 est assez proche de L2. Et cela interdit de discerner des classes de langues équivalentes en proximité. La propriété de distance thématique entre langues est singulière et propre à chaque langue. Ainsi, par exemple, la Toile en hébreu

et la Toile francophone sont toutes les deux assez proches de la Toile anglophone, en ce qui concerne les thématiques, cependant elles sont toutes les deux assez éloignées l'une de l'autre.

Cette représentation permet de constater :

- La forte proximité des web en espagnol et en anglais.
- La forte proximité du web en suédois avec les web danois, hollandais, norvégien et finlandais.
- La forte singularité de l'hindi qui n'est proche d'aucune autre langue.
- Le web chinois assez éloigné du reste à l'exception de l'hindi, du russe et du turc.
- La singularité du web allemand qui est seulement assez proche des web finlandais, hollandais, italien et espagnol.

5.2 Impact économique

Pour chaque langue, un index, qui expime le poids économique moyen des sites pour chaque langue, a été calculé selon la méthode indiqué dans le chapitre méthodologie. Pour rechercher une certaine cohérence, la mesure a également été réalisée pour le domaine Internet du pays correspondant le mieux à la langue. Pour certaine langue, comme le portugais, les mesures sont réalisées pour deux pays, dans ce cas Brésil et Portugal. Les résultats sont présentés dans le tableau suivant trié en poids décroissant. Le français se classe en 4^{ième} position. Le hollandais est un solide vainqueur, et, au fond du classement, un surprenant coréen derrière chinois et marathi. Il faut souligner que l'étude des biais linguistiques de la base de données, réalisée au chapitre 5.3, a montré que les biais favorisaient justement le hollandais le japonais et le polonais ainsi que l'allemand et l'italien. Il n'est pas impossible que des effets de ce biais affectent ce résultat qui place cependant le français en très bonne position sur un critère essentiel, celui de l'impact économique.

Tableau 12 : Impact économique

	Poids Langue	Poids Domaine1	Poids Domaine2	MOYENNE
MONDIAL				**18,98**
Hollandais	21,38	21,23		21,30
Japonais	19,48	21,35		20,41
Polish	20,33	20,18		20,26
Français	20,13	19,65		19,89
Italien	19,86	19,74		19,80
Portugais	19,03	19,11	20,27	19,47
Allemand	19,64	19,10		19,37
Ourdou	19,49	19,15		19,32
Ukrainien	19,09	19,51		19,30
Malaisien	19,23	19,22	19,34	19,27
Espagnol	19,06	19,36		19,21
Thaïlandais	18,37	20,02		19,20
Arabe	18,20	19,54		18,87
Turc	18,33	19,26		18,80
Vietnamien	18,69	18,87		18,78

Anglais	19,17	17,75	19,31	18,75
Perse	18,53	17,83		18,18
Kannada	18,43	17,90		18,17
Tamil	18,41	17,90		18,16
Russe	18,29	18,00		18,15
Gujarati	18,23	17,90		18,07
Telugu	17,87	17,90		17,89
Swahili	17,84	17,90		17,87
Javanais	16,24	19,22		17,73
Hindi	16,87	17,90		17,39
Chinois	17,00	15,49		16,25
Marathi	14,29	17,90		16,09
Coréen	16,17	15,95		16,06

5.3 Degré de confiance pour commerce électronique

Le tableau suivant classe en ordre décroissant les langues dont les sites de commerce électronique reçoivent les meilleures notes de confiance. Il faut noter que les écarts sont faibles et donc les biais linguistiques susceptibles d'intervenir. Là aussi la Toile francophone obtient de bons résultats, en haut du classement, avec des écarts si faibles qu'ils sont probablement susceptibles aux différents biais.

Tableau 13 : Confiance envers commerce électronique

	Poids Langue	**Poids Domaine1**	**Poids Domaine2**	**MOYENNE**
MONDIAL				**94,38**
Italien	97,64	97,23		97,44
Allemand	97,81	96,68		97,25
Hollandais	97,26	95,97		96,61
Portugais	96,72	95,93	96,34	96,33
Espagnol	96,56	95,93		96,24
Polish	96,89	95,46		96,17
Français	96,29	95,89		96,09
Ukrainien	96,70	95,36		96,03
Turc	95,97	95,63		95,80
Coréen	95,59	95,88		95,74
Japonais	95,15	96,08		95,61
Thaïlandais	93,98	96,05		95,01
Arabe	94,60	94,56		94,58
Anglais	95,97	92,83		94,40
Malaisien	91,55	93,94	96,27	93,92
Tamil	93,71	93,84		93,78
Kannada	93,69	93,84		93,76

Gujarati	93,42	93,84		93,63
Russe	94,25	92,81		93,53
Vietnamien	93,26	93,08		93,17
Telugu	92,48	93,84		93,16
Hindi	92,32	93,84		93,08
Marathi	91,66	93,84		92,75
Perse	94,98	89,88		92,43
Swahili	91,13	92,63		91,88
Ourdou	90,36	93,30		91,83
Chinois	94,14	88,70		91,42
Javanais	83,88	93,94		88,91

5.4 Nombre moyen de pages par site

Tableau 14 : Nombre de pages par site

LANGUES	PAGES	%SITES < 25 pages
MONDIAL	**31,90**	**68%**
Javanais	18,27	79%
Allemand	21,96	65%
Anglais	22,33	72%
Marathi	26,58	69%
Hollandais	26,62	67%
Portugais	28,36	69%
Italien	30,17	60%
Espagnol	30,49	66%
Français	31,51	66%
Malaisien	32,06	65%
Thaïlandais	33,19	67%
Turc	37,12	57%
Japonais	38,10	57%
Polish	38,94	59%
Arabe	40,42	58%
Hindi	47,19	46%
Swahili	48,28	55%
Russe	49,87	53%
Coréen	53,90	46%
Ukrainien	58,79	48%
Telugu	61,86	43%
Tamil	64,16	44%
Vietnamien	66,33	33%
Chinois	66,64	34%
Gujarati	67,58	41%
Perse	67,87	35%

Kannada	73,19	31%
Ourdou	88,24	30%

Il n'est pas évident d'interpréter ce résultat. Au vu des langues en bas du classement, avec deux exceptions, le javanais et le marathi, il semblerait qu'un nombre moyen de pages par site élevé est le signe d'un écosystème concentré sur les sites importants, alors qu'une moyenne faible est le signe d'un écosystème dynamique avec de nombreux sites de petite taille. Le pourcentage de sites avec moins de 25 pages est là pour conforter cette hypothèse. Les deux exceptions sont peut-être des écosystèmes dans une étape débutante de prototype avec rareté de sites importants, ou bien encore un biais dû à une couverture statistiquement trop faible pour être représentative. Dans tous les cas, il y a peu d'enseignements à tirer de ce paramètre pour lequel les langues européennes dont le français ont des résultats très comparables avec toutefois une avance marquée pour l'anglais et l'allemand qui aurait donc une Toile tissée plus fine que le reste des langues

En marge de sa définition, ce paramètre donne l'occasion de faire une comparaison entre la représentation linguistique de DataProvider et celle du modèle d'OBDILCI. Nous savons que DataProvider classe les langues, comme W3Techs, sans tenir compte du multilinguisme possible des sites web, et cela explique des pourcentages affichés de l'anglais de plus du double de ceux d'OBDILCI[8]. Nous soupçonnons aussi un biais naturel, étant donné la niche d'affaire de la compagnie, envers son pays de résidence, les Pays-Bas et son voisin linguistique l'Allemagne, qui provoque des pourcentages du hollandais et de l'allemand favorisés par un plus fort pourcentage de sites inclus dans la base de données, proche de l'exhaustivité. Nous allons vérifier ces hypothèses par des calculs.

A partir du nombre de pages moyen par langue, nous pouvons construire les indicateurs linguistiques du nombre moyen de pages par locuteur et nombre moyen de pages par locuteur connecté. Cela permet finalement calculer les mêmes indicateurs que ceux du modèle d'OBDILCI :

PV (présence virtuelle) = pourcentage de pages divisé par pourcentage de locuteurs
PC (productivité des contenus) = pourcentage de pages divisé par pourcentage de locuteurs connectés)

Cette comparaison permettra de confirmer ou non les hypothèses sur les biais linguistiques de la base de donnéees. Cette élaboration est faite en utilisant les données d'Ethnologue pour les locuteurs et une élaboration d'OBDILCI à partir des donnéees de l'UIT, pour les pourcentages de locuteurs connectés.

Une définition équivalente de PV et PC, qui transforme l'approche par pourcentage en approche directe est, pour chaque langue l :

PV(l) = Nombre de pages par locuteur (l) / nombre de pages par locuteur pour toutes les langues

[8] Voir « Est-il vrai que plus de la moitié du contenu Web est en anglais ? Pas si l'on prend en compte le multilinguisme ! », version en anglais publiée dans Forum of Linguistic Studies| Volume 06 | Numéro 05 | Novembre 2024.

PC(l) = Nombre de pages par locuteur connecté (l) / nombre de pages par locuteur connecté pour toutes les langues.

A partir de cette équation, on obtient les résultats suivants, où PV-C est l'indicateur PV calculé sur les donnéees de DataProvider alors que PV est la valeur du modèle d'OBDILCI ; de même pour PC. La dernière colonne indique le rapport entre les valeurs calculées et les valeurs du modèle OBDILCI. Un nombre très supérieur à 1 indique un biais d'autant plus fort dans la sélection pour la base de donnéees de DataProvider et un nombre très inférieur à 1 indique une couverture linguistique extrêmement faible dans la base de DataProvider.

Tableau 15 : Comparaison indicateurs avec OBDILCI

	Page/ LOC	PV-C	PV	PV-C/ PV
Hollandais	**1,227**	4,64	**1,66**	2,790
Polish	**0,595**	2,25	**1,44**	1,559
Allemand	**0,674**	2,55	**1,72**	1,482
Anglais	**0,546**	2,07	**1,44**	1,438
Japonais	**0,701**	2,65	**1,89**	1,402
Italien	**0,556**	2,10	**1,59**	1,323
Coréen	**0,338**	1,28	**1,15**	1,110
Français	**0,318**	1,20	**1,17**	1,027
Perse	**0,224**	0,85	**0,86**	0,987
Vietnamien	**0,302**	1,14	**1,19**	0,960
Turc	**0,298**	1,13	**1,34**	0,841
Russe	**0,326**	1,23	**1,57**	0,784
Ukrainien	**0,237**	0,90	**1,20**	0,748
Portugais	**0,216**	0,82	**1,28**	0,639
Malaisien	**0,124**	0,47	**0,89**	0,530
Thaïlandais	**0,142**	0,54	**1,18**	0,457
Espagnol	**0,158**	0,60	**1,48**	0,404
Chinois	**0,140**	0,53	**1,32**	0,401
Arabe	**0,027**	0,10	**0,89**	0,115
Marathi	**0,010**	0,04	**0,63**	0,060
Hindi	**0,008**	0,03	**0,67**	0,045
Tamil	**0,006**	0,02	**0,66**	0,037
Ourdou	**0,003**	0,01	**0,41**	0,025
Kannada	**0,004**	0,01	**0,62**	0,024
Gujarati	**0,004**	0,01	**0,64**	0,022
Swahili	**0,002**	0,01	**0,37**	0,020
Telugu	**0,003**	0,01	**0,64**	0,018
Javanais	**0,001**	0,00	**0,81**	0,004
TOTAL	**0,265**			

	Pages/ Locc	PC-C	PC	PC-C/ PC
Hollandais	**2,858**	6,31	**1,11**	**5,682**
Allemand	**1,511**	3,34	**1,20**	**2,792**
Polish	**1,257**	2,78	**1,06**	**2,610**
Italien	**1,181**	2,61	**1,16**	**2,243**
Japonais	**1,447**	3,20	**1,43**	**2,238**
Anglais	**0,944**	2,09	**1,29**	1,611
Russe	**0,702**	1,55	**1,14**	1,365
Turc	**0,618**	1,36	**1,01**	1,354
Vietnamien	**0,575**	1,27	**0,97**	1,307
Ukrainien	**0,468**	1,03	**0,95**	1,093
Coréen	**0,541**	1,20	**1,12**	1,069
Français	**0,519**	1,15	**1,12**	1,024
Perse	**0,351**	0,77	**0,85**	0,907
Portugais	**0,406**	0,90	**1,06**	0,846
Thaïlandais	**0,309**	0,68	**0,84**	0,808
Malaisien	**0,217**	0,48	**0,79**	0,603
Espagnol	**0,312**	0,69	**1,16**	0,594
Chinois	**0,266**	0,59	**1,08**	0,542
Arabe	**0,045**	0,10	**0,84**	0,117
Marathi	**0,012**	0,03	**0,83**	0,031
Hindi	**0,009**	0,02	**0,88**	0,024
Tamil	**0,008**	0,02	**0,85**	0,021
Kannada	**0,005**	0,01	**0,83**	0,012
Gujarati	**0,004**	0,01	**0,85**	0,012
Telugu	**0,004**	0,01	**0,84**	0,009
Ourdou	**0,003**	0,01	**0,70**	0,008
Swahili	**0,002**	0,00	**0,73**	0,005
Javanais	**0,001**	0,00	**0,75**	0,004
TOTAL	**0,453**			

La comparaison confirme et étend les premières impressions basées sur la simple comparaison du classement des langues de DataProvider :

- Certains pays bénéficient d'un biais très positif en termes de sites traités dans la base de données : Pays-Bas (pays siège de l'entreprise), Allemagne, Pologne, Italie et Japon.
- Un autre groupe bénéficie d'un biais positif important : pays anglophones, Russie et Turquie.
- Un groupe souffre d'un biais négatif modéré : Malais, pays hispanophones, Chine.
- Un dernier groupe reçoit un traitement extrêmement marginal, en particulier les langues de l'Inde et d'Afrique et dans une moindre mesure l'arabe.

Le cas du français est particulier : d'un côté le français est très bien couvert dans les pays du Nord, de l'autre côté il est marginalisé dans les pays du Sud et en particulier en Afrique. Cette affirmation résulte de la possibilité d'interroger la base par région et de croiser cette interrogation par langue. Ainsi, le décompte de sites francophones en Afrique est de seulement 369 sur les 3 343 082 sites en français, alors que pour l'Amérique du Sud il est de 1483. Par contre, et pour compenser, il semble que la France et les pays du nord avec forte population francophone fassent partie des pays traités avec un biais positif.

Ces considérations doivent être prises en compte dans l'ensemble des résultats présentés comme des biais affectant les langues, biais positifs ou négatifs, biais marginaux ou majuscules, selon les langues.

5.5 Nombre moyen de liens entrants par site

Le nombre de liens entrants est un paramètre essentiel dans la célébrité d'un site et indirectement influe dans les classements dans les réponses de moteurs de de recherche. Les résultats montrent des usages qui différent fortement entre les langues. Pour certaines, celles d'une certaine manière du web occidental, l'usage de plus en plus prégnant, est de ne pas offrir facilement ces liens entrants. Dans d'autres cas, surtout celui de la Chine, il semble y avoir une grande générosité à offrir des liens. Pour les cas de l'ourdou et du malais et du javanais, les différences sont tellement fortes entre les résultats par langue et ceux par domaine pays, qu'ils font suspecter des biais importants dans la détection des langues et il est préférable de ne pas en tenir compte.

Tableau 16 : Liens entrants

LIENS ENTRANTS	/Langue	/Domaine1	/Domaine2	MOY	SANS LIEN %
MONDIAL				**5,97**	**68,3**
Chinois	**24,92**	**28,33**		26,62	60,8
~~Ourdou~~	~~18,11~~	~~3,61~~	=	~~10,86~~	~~67,5~~
~~Malais~~	~~15,67~~	~~**9,85**~~	**4,29**	~~9,93~~	~~72,7~~
Hollandais	**8,32**	**8,50**		8,41	54,0
Turc	6,74	**8,49**		7,61	71,7
Allemand	6,75	**7,02**		6,88	57,8
Thaïlandais	6,23	**7,47**		6,85	74,8
~~Javanais~~	~~3,09~~	~~**9,85**~~	=	~~6,47~~	~~83,6~~

Vietnamien	6,96	**5,22**		6,09	70,5
Polonais	5,70	**5,56**		5,63	61,7
Japonais	5,64	**5,43**		5,54	58,6
Français	5,26	**4,88**		5,07	61,1
Anglais	4,84	**4,02**	**5,13**	4,66	71,3
Italien	4,47	**4,83**		4,65	61,0
Persan	4,57	**4,46**		4,51	73,1
Ukrainien	4,20	**4,52**		4,36	69,4
Russe	4,01	**4,26**		4,13	72,2
Espagnol	3,15	**4,88**		4,02	71,1
Coréen	3,95	**4,00**		3,97	76,4
Marathi	5,37	**2,37**		3,87	81,2
Gujarati	4,89	**2,37**		3,63	77,4
Portugais	**3,08**	**2,75**	**4,95**	3,59	75,8
Tamil	4,48	**2,37**		3,42	74,3
Arabe	3,34	**3,43**		3,39	80,4
Kannada	4,04	**2,37**		3,20	76,4
Hindi	3,75	**2,37**		3,06	79,1
Telugu	3,43	**2,37**		2,90	81,6
Swahili	3,73	**2,04**		2,89	78,3

5.6 Orientation d'affaires des sites

Ce paramètre permet d'établir, dans l'échantillon choisi, le pourcentage de sites ayant vocation d'affaires et à l'intérieur de ce groupe, les pourcentages relatifs de sites à vocation B2C (vers le consommateur) ou B2B (entre négoces), ce dernier représentant un état plus évolué.

Les deux tableaux présentent les même donnéees et sont seulement différenciés par le tri, par pourcentage décroissant de sites d'affaires pour celui de gauche, par pourcentage décroissant de sites B2B, pour celui de droite.

Les résultats du pourcentage de sites à vocation d'affaires peuvent être rapprochés de l'étude du paramètre « *thématique* », en conscience que pour la plupart des thématiques il est possible ou non d'avoir une vocation d'affaires.

Ainsi, les sites en chinois sont au plus bas dans la catégorie affaires mais dans ce groupe ils sont essentiellement B2B, avec une note très basse pour le B2C. Le français se classe très proche de la plupart des langues européennes, à l'exception de l'espagnol et du portugais pour lesquelles il faut rechercher l'explication du côté de l'Amérique latine.

Nous avons écarté deux paramètres (« heartbeat », qui renseigne sur le degré d'actualisation des sites et « Website age » sur l'ancienneté du site), à cause de doutes sérieux sur leur validité et nous avons hésité à en faire de même pour ce paramètre, à interpréter donc avec prudence.

En ce qui concerne, le web francophone, il se place dans la moyenne en termes d'orientation vers les affaires et une orientation forte dans le B2C, avec une performance moyenne haute dans le B2B.

Tableau 17 : Classement des langues pour les critères affaires, b2b et b2c

LANGUE	AFFAIRES	B2B	B2C
Grec	**93,6%**	10,32%	38,70%
Polonais	**92,7%**	14,32%	33,42%
Hébreu	**91,8%**	7,05%	23,02%
Perse	**91,5%**	3,03%	12,02%
Arabe	**90,7%**	5,30%	12,36%
Espagnol	**89,7%**	17,95%	36,97%
Italien	**89,0%**	12,42%	40,51%
Estonien	**88,0%**	5,17%	42,29%
Allemand	**87,4%**	9,96%	45,21%
Turc	**87,0%**	14,46%	22,83%
Portugais	**86,6%**	15,65%	27,41%
Vietnamien	**86,5%**	0,93%	41,90%
Danois	**84,6%**	7,92%	40,37%
Français	**84,4%**	9,86%	42,33%
Russe	**84,4%**	11,76%	16,09%
Finlandais	**84,2%**	9,06%	36,80%
Hollandais	**83,3%**	10,85%	47,06%
Suédois	**82,4%**	10,69%	29,65%
Coréen	**80,8%**	4,27%	26,44%
Norvégien	**80,6%**	7,44%	29,79%
Anglais	**74,8%**	11,29%	36,07%
Malais	**64,9%**	2,84%	21,08%
Hindi	**60,6%**	1,43%	11,40%
Japonais	**52,6%**	8,65%	22,99%
Thaïlandais	**40,7%**	3,19%	17,61%
Chinois	**32,5%**	11,76%	2,35%

LANGUE	AFFAIRES	B2B	B2C
Espagnol	89,7%	**17,95%**	36,97%
Portugais	86,6%	**15,65%**	27,41%
Turc	87,0%	**14,46%**	22,83%
Polonais	92,7%	**14,32%**	33,42%
Italien	89,0%	**12,42%**	40,51%
Russe	84,4%	**11,76%**	16,09%
Chinois	32,5%	**11,76%**	2,35%
Anglais	74,8%	**11,29%**	36,07%
Hollandais	83,3%	**10,85%**	47,06%
Suédois	82,4%	**10,69%**	29,65%
Grec	93,6%	**10,32%**	38,70%
Allemand	87,4%	**9,96%**	45,21%
Français	84,4%	**9,86%**	42,33%
Finlandais	84,2%	**9,06%**	36,80%
Japonais	52,6%	**8,65%**	22,99%
Danois	84,6%	**7,92%**	40,37%
Norvégien	80,6%	**7,44%**	29,79%
Hébreu	91,8%	**7,05%**	23,02%
Arabe	90,7%	**5,30%**	12,36%
Estonien	88,0%	**5,17%**	42,29%
Coréen	80,8%	**4,27%**	26,44%
Thaïlandais	40,7%	**3,19%**	17,61%
Perse	91,5%	**3,03%**	12,02%
Malais	64,9%	**2,84%**	21,08%
Hindi	60,6%	**1,43%**	11,40%
Vietnamien	86,5%	**0,93%**	41,90%

6 - CONCLUSIONS

Les thèmes les plus abondamment traités dans la Toile francophone, par rapport aux 18 autres langues mentionnées, sont dans l'ordre : *Gastronomie, Voyage et tourisme, Nature, Art et design, Gouvernement & Société, Soins de beauté et personnels, Telecom, Science, Mode et Médias.*

Ce n'est pas une surprise, même si une meilleure place pour la culture (représentée approximativement par « art et design » aurait été attendue).

Certains des thèmes les moins abondamment traités dans la Toile francophone sont peut-être ceux qui peuvent apporter les enseignements les plus intéressants : *Finance, Logiciels et données, Education, Affaires et carrières et Transport*.

Les trois résultats les plus bas interrogent.

La dernière place pour *finance* traduit-elle vraiment un manque d'intérêt culturel ou un plutôt un écosystème trop centralisé et donc avec un nombre réduit d'acteurs ?

Au moment où les *données* deviennent, avec l'intelligence artificielle, le centre d'attention prioritaire de toutes les économies, faut-il se préoccuper, là aussi d'un positionnement en queue de classement sur ce thème ? Est-ce le signe de l'absence d'une économie décentralisée et d'une dynamique créatrice trop faible ?

Pourquoi la Toile francophone comporte-t-elle une attention à *l'éducation* si limitée en quantité de sites ? Quelle est la signification de ce phénomène ? Est-ce que cela est compensée par la qualité et le dynamisme des sites présents, moins nombreux mais plus puissants ?

Cette étude, avec les biais inévitables et non maitrisables qu'elle comporte, appelle à regarder de plus près les résultats extrêmes, moins susceptibles d'être affecter par les biais. Elle pose plus de questions qu'elle n'apporte de réponses. Peut-être permet-elle de poser certaines questions importantes.

Pour le reste des paramètres, il n'y a rien qui permet de lever des alarmes, au contraire, il semble que la Toile francophone se porte bien au niveau de son impact économique et ne montre pas de faiblesse notable dans les paramètres étudiés.

6- ANNEXES

6.1 CLASSEMENTS DES LANGUES PAR THEME

La consultation de ces tableaux permet d'affiner les constats relatifs à chaque thème particulier et, pour les personnes intéressées, de se concentrer sur la langue de son choix pour établir la signature thématique de son Web.

L'observation des tableaux montre :

- que la Toile italienne est très semblable à la Toile francophone ;
- que le web en Hindi paye ses trois premières places mais surtout son avance géante en « divertissement » en prenant la dernière place dans un grand nombre de thème ;

Health	
M.POND.	7,24%
LANG.	%
German	11,87%
Japanese	10,63%
Danish	10,43%
Hebrew	8,91%
Spanish	8,16%
Dutch	7,90%
Finnish	7,12%
Portuguese	6,99%
Swedish	6,73%
French	6,67%
Italian	6,63%
English	6,23%
Polish	6,23%
Turkish	6,23%
Korean	6,23%
Norwegian	5,37%
Russian	3,25%
Chinese	2,88%
Hindi	0,63%

Gov. & Society	
M.POND.	2,95%
LANG.	%
Hindi	7,60%
Russian	6,23%
Dutch	4,13%
French	3,98%
Italian	3,55%
Chinese	3,28%
Finnish	3,24%
German	3,06%
Japanese	2,80%
Swedish	2,69%
Norwegian	2,56%
Danish	2,51%
Korean	2,51%
Portuguese	2,49%
Spanish	2,48%
Turkish	2,21%
Hebrew	2,11%
English	2,07%
Polish	2,02%

Travel & Society	
M.POND.	4,67%
LANG.	%
Italian	9,09%
French	8,16%
Norwegian	7,58%
German	7,40%
Korean	6,28%
Polish	5,92%
Swedish	5,32%
Finnish	5,14%
Japanese	4,92%
Dutch	4,40%
Spanish	4,03%
Russian	3,68%
Hebrew	3,68%
Turkish	3,65%
English	3,19%
Danish	2,98%
Portuguese	2,58%
Chinese	1,14%
Hindi	0,51%

Food & Drink	
M.POND.	6,67%
LANG.	%
Italian	10,81%
French	10,49%
Korean	8,82%
German	8,30%
Japanese	8,26%
Dutch	7,85%
Spanish	7,42%
Danish	7,17%
Swedish	6,11%
English	5,83%
Hebrew	5,37%
Finnish	5,28%
Norwegian	5,24%
Polish	5,05%
Turkish	4,90%
Portuguese	4,38%
Russian	2,99%
Chinese	2,81%
Hindi	1,49%

Telecom	
MOY.POND.	1,53%
LANG.	%
English	1,66%
Portuguese	0,80%
Hebrew	0,59%
Russian	0,53%
French	0,46%
Spanish	0,29%
Dutch	0,28%
Polish	0,25%
Turkish	0,24%
Italian	0,20%
Korean	0,17%
Japanese	0,14%
Norwegian	0,14%
Hindi	0,10%
Swedish	0,10%
Danish	0,08%
Chinese	0,07%
German	0,07%
Finnish	0,05%

A noter, l'avance confortable pour l'italien et le français dans le thème *gastronomie*, de l'allemand dans le thème *santé*, et le cavalier seul de l'anglais dans le thème des *télécommunications* avec plus du double de pourcentage de chacun des suivants.

Religious		Financial		Adult		Fashion		Science	
M.POND.	**2,52%**	**M.POND.**	**4,05%**	**M.POND.**	**4,00%**	**M.POND.**	4,99%	**M.POND..**	1,08%
LANG.	**%**	**LANG.**	**%**	**LANG.**	**%**	**LANG.**	%	**LANG.**	%
Hindi	**5,19%**	**Portuguese**	**8,92%**	**Chinese**	**8,03%**	**Korean**	**22,01%**	**Russian**	**1,30%**
Hebrew	**3,58%**	**Norwegian**	**5,74%**	Turkish	2,51%	**Portuguese**	**6,39%**	**English**	**1,17%**
Korean	**3,05%**	**Polish**	**4,95%**	Japanese	2,38%	**Spanish**	**5,34%**	**Portuguese**	**1,16%**
English	**2,79%**	**Hebrew**	**4,87%**	Hindi	1,45%	English	4,69%	Italian	0,72%
German	2,39%	**Finnish**	**4,51%**	Russian	1,16%	French	4,65%	French	0,69%
Italian	1,98%	**German**	**4,15%**	Hebrew	0,94%	Italian	3,90%	Chinese	0,66%
Polish	1,95%	**Japanese**	**4,09%**	Dutch	0,78%	Hebrew	3,88%	Spanish	0,63%
Norwegian	1,70%	Turkish	3,77%	Norwegian	0,63%	Swedish	3,21%	Polish	0,56%
Spanish	1,67%	Russian	3,72%	Swedish	0,59%	Turkish	3,05%	Hebrew	0,41%
Finnish	1,58%	English	3,71%	Danish	0,55%	Norwegian	3,01%	Japanese	0,41%
Dutch	1,45%	Swedish	3,16%	German	0,51%	Japanese	2,89%	Turkish	0,38%
Portuguese	1,38%	Spanish	2,87%	Finnish	0,47%	Danish	2,87%	Norwegian	0,29%
Turkish	1,29%	Hindi	2,71%	French	0,46%	Polish	2,86%	Korean	0,28%
French	1,24%	Korean	2,70%	English	0,44%	Finnish	2,58%	Hindi	0,26%
Danish	1,15%	Italian	2,67%	Korean	0,34%	German	2,23%	Finnish	0,23%
Swedish	1,06%	Dutch	2,49%	Spanish	0,29%	Dutch	2,17%	Danish	0,21%
Russian	0,77%	Chinese	2,45%	Portuguese	0,27%	Russian	1,81%	German	0,20%
Japanese	0,74%	Danish	2,07%	Italian	0,23%	Hindi	1,21%	Swedish	0,17%
Chinese	0,40%	French	1,76%	Polish	0,22%	Chinese	1,03%	Dutch	0,12%

A noter la dernière place isolée du français dans le thème *finance* et l'avance confortable du portugais dans ce thème qui pointe vers le Brésil. A noter également, les 4 premières places pour le thème *religion* de l'hindi, l'hébreu, le coréen et l'anglais, le score différent du norvégien par rapport aux autres pays nordiques et une place du turc proche de celle du français.

Education		Home & Garden		Construct.		Transport		Art & design	
M.POND.	4,10%	**M.POND.**	**5,95%**	**M.POND.**	**11,25%**	**M.POND.**	**4,08%**	**M.POND.**	**8,88%**
LANG.	%	**LANG.**	**%**	**LANG.**	%	**LANG.**	%	**LANG.**	Count/L
Polish	**8,02%**	**Korean**	**18,60%**	**Turkish**	**23,30%**	**Turkish**	**8,07%**	**German**	**11,24%**
Japanese	**6,77%**	**Hebrew**	**9,23%**	**Finnish**	**20,78%**	**Russian**	**7,29%**	**Italian**	**9,87%**
Portuguese	**6,40%**	**Dutch**	**8,25%**	**Swedish**	**19,13%**	**Finnish**	**7,13%**	**French**	**9,42%**
Hebrew	**6,24%**	**Danish**	**8,15%**	**Russian**	**19,05%**	**Polish**	**5,98%**	**English**	**9,12%**
Hindi	**5,19%**	**Polish**	**7,68%**	**Polish**	**16,23%**	**Portuguese**	**5,39%**	**Spanish**	**9,03%**
Turkish	**5,14%**	**Russian**	**7,48%**	**Norwegian**	**14,63%**	**Dutch**	**4,98%**	Dutch	6,66%
Spanish	**4,65%**	**Turkish**	**7,20%**	**Japanese**	**13,70%**	**Spanish**	**4,88%**	Hebrew	6,65%
German	**4,30%**	**French**	**7,18%**	**French**	**13,31%**	**Swedish**	**4,84%**	Swedish	5,19%
Swedish	**4,14%**	**Spanish**	**6,66%**	**Danish**	**12,16%**	**German**	**4,67%**	Finnish	4,95%
Russian	4,01%	**Italian**	**6,34%**	**Dutch**	**11,66%**	**Danish**	**4,59%**	Korean	4,89%
Danish	3,90%	**Swedish**	**6,16%**	Hebrew	10,31%	**Italian**	**4,53%**	Danish	4,80%
Chinese	3,58%	Norwegian	5,65%	German	9,93%	**Norwegian**	**4,47%**	Norwegian	4,58%
Dutch	3,55%	English	5,29%	Italian	9,65%	**French**	**4,43%**	Polish	4,45%
English	3,41%	Chinese	4,71%	English	9,24%	Japanese	3,29%	Japanese	3,00%
Finnish	3,08%	Finnish	4,54%	Spanish	8,64%	English	2,74%	Portuguese	2,14%
French	2,97%	Portuguese	4,53%	Korean	8,14%	Hebrew	2,65%	Turkish	1,98%
Italian	2,59%	German	4,02%	Portuguese	7,52%	Chinese	1,27%	Russian	1,89%
Norwegian	2,39%	Japanese	3,04%	Chinese	6,43%	Korean	0,89%	Chinese	1,35%
Korean	1,96%	Hindi	1,98%	Hindi	0,42%	Hindi	0,30%	Hindi	0,66%

Le thème *éducation* mériterait une analyse plus poussée avec d'autres moyens. Il n'y a pas que la place du français qui surprend mais aussi celle des 7 dernières langues autant que celle du polonais en confortable première place. A noter l'avance importante de l'allemand pour le thème art et design.

Beauty & Pers. care	
M.POND.	**7,34%**
LANG.	**%**
Japanese	**10,69%**
Danish	**9,19%**
Finnish	**8,15%**
French	**8,08%**
Swedish	**7,96%**
English	**7,51%**
Portuguese	6,99%
Spanish	6,39%
Polish	6,18%
German	5,71%
Hebrew	5,68%
Dutch	5,63%
Norwegian	5,57%
Korean	5,17%
Italian	4,31%
Turkish	3,09%
Russian	2,75%
Hindi	1,56%
Chinese	0,30%

Software & Data	
M.POND.	**8,80%**
LANG.	**%**
Russian	**12,33%**
Chinese	**11,96%**
Norwegian	**10,88%**
Swedish	**10,35%**
Spanish	**9,60%**
English	8,68%
German	8,47%
Portuguese	7,99%
Dutch	7,81%
Korean	6,82%
Finnish	5,45%
Hebrew	4,89%
Hindi	4,86%
Danish	4,71%
Turkish	4,68%
Italian	4,61%
French	4,46%
Japanese	4,34%
Polish	4,25%

Nature	
M.POND.	**1,28%**
LANG.	**%**
English	**1,41%**
Norwegian	**1,35%**
French	**1,33%**
Spanish	1,16%
Japanese	1,09%
Finnish	1,01%
Dutch	0,94%
Polish	0,93%
Danish	0,89%
German	0,88%
Hebrew	0,75%
Portuguese	0,69%
Italian	0,59%
Russian	0,50%
Swedish	0,43%
Turkish	0,27%
Korean	0,24%
Chinese	0,23%
Hindi	0,13%

Business & Careers	
M.POND.	**1,43%**
LANG.	**%**
Norwegian	**3,72%**
Swedish	**3,26%**
Hindi	**3,23%**
Portuguese	**2,76%**
Danish	**2,60%**
Dutch	**2,43%**
Italian	**2,13%**
Korean	**2,01%**
Spanish	1,48%
Polish	1,43%
Finnish	1,40%
German	1,29%
French	1,24%
English	1,00%
Hebrew	0,92%
Russian	0,64%
Japanese	0,50%
Turkish	0,25%
Chinese	0,03%

Hardware & Electr .	
M.POND.	**2,79%**
LANG.	**%**
Chinese	**5,22%**
Portuguese	**4,69%**
Norwegian	**4,23%**
Danish	**3,88%**
Korean	**3,37%**
Hebrew	**3,36%**
Turkish	**3,31%**
Polish	2,71%
Spanish	2,62%
French	2,50%
Russian	2,48%
Italian	2,38%
English	2,38%
Swedish	1,89%
Dutch	1,67%
Finnish	1,59%
Japanese	1,55%
German	1,11%
Hindi	1,03%

La compagnie du japonais avec le français en fin de classement pour la catégorie *software et données* surprend, encore un thème qui mériterait une analyse complémentaire. Le score epsilonesque du chinois dans le secteur *affaires et carrières* de quoi est-il le nom ?

Entertainment	
M.POND.	**15,64%**
LANG.	**%**
Hindi	**77,97%**
Chinese	**35,08%**
Japanese	**21,18%**
German	14,40%
Portuguese	12,71%
English	12,43%
Norwegian	12,13%
French	12,03%
Korean	11,78%
Dutch	10,95%
Danish	10,82%
Italian	9,63%
Spanish	9,24%
Turkish	9,04%
Russian	8,77%
Finnish	8,28%
Swedish	6,79%
Polish	6,18%
Hebrew	5,96%

Juridic	
M.POND.	**2,16%**
LANG.	%
Hebrew	**6,81%**
Portuguese	**4,30%**
Polish	**3,04%**
German	**3,03%**
Italian	**2,57%**
Spanish	**2,52%**
Russian	**2,24%**
Japanese	1,97%
Korean	1,72%
Turkish	1,67%
English	1,66%
French	1,64%
Norwegian	1,01%
Finnish	0,93%
Chinese	0,78%
Dutch	0,76%
Swedish	0,62%
Danish	0,58%
Hindi	0,06%

Marketing	
M.POND.	**4,35%**
LANG.	%
Portuguese	**6,33%**
Hebrew	**6,31%**
English	**4,57%**
German	**4,34%**
Chinese	**3,93%**
Spanish	**3,85%**
Korean	**3,71%**
Danish	**3,65%**
Turkish	**3,52%**
Italian	**3,25%**
French	**3,24%**
Finnish	**2,99%**
Dutch	**2,73%**
Polish	**2,61%**
Norwegian	**2,23%**
Swedish	**2,19%**
Japanese	**2,00%**
Russian	**1,36%**
Hindi	**0,89%**

Medias	
M.POND.	**5,02%**
LANG.	%
English	**5,61%**
Hebrew	3,83%
German	3,27%
Italian	3,03%
French	2,91%
Dutch	2,32%
Norwegian	2,28%
Spanish	2,12%
Russian	1,94%
Finnish	1,82%
Swedish	1,72%
Danish	1,51%
Polish	1,25%
Portuguese	1,11%
Turkish	1,05%
Korean	0,63%
Chinese	0,44%
Japanese	0,41%
Hindi	0,35%

La part disproportionnée du *divertissement* dans les trois pays asiatiques de tête est frappante. Quelle serait la raison d'une solide première place de l'hébreu dans le secteur *juridique* ?

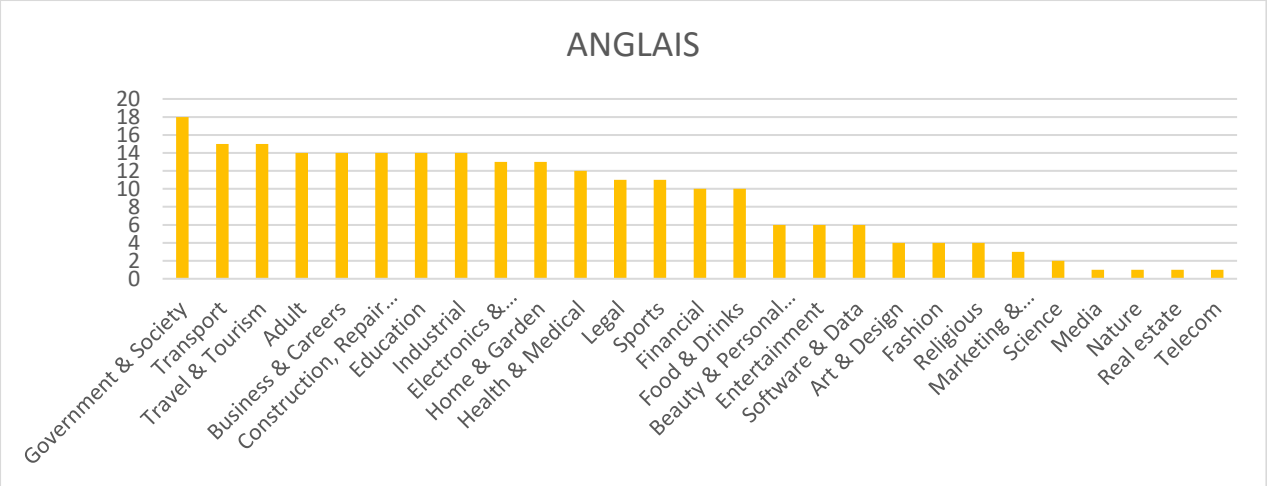
Real Estate	
M.POND.	**3,45%**
LANG.	%
English	**3,75%**
Hebrew	**3,68%**
Portuguese	**3,58%**
Spanish	**3,48%**
Italian	**3,43%**
Dutch	3,14%
German	2,87%
Korean	2,41%
French	2,15%
Russian	2,04%
Japanese	1,74%
Polish	1,71%
Swedish	1,69%
Turkish	1,59%
Danish	1,13%
Norwegian	1,07%
Chinese	0,86%
Finnish	0,80%
Hindi	0,19%

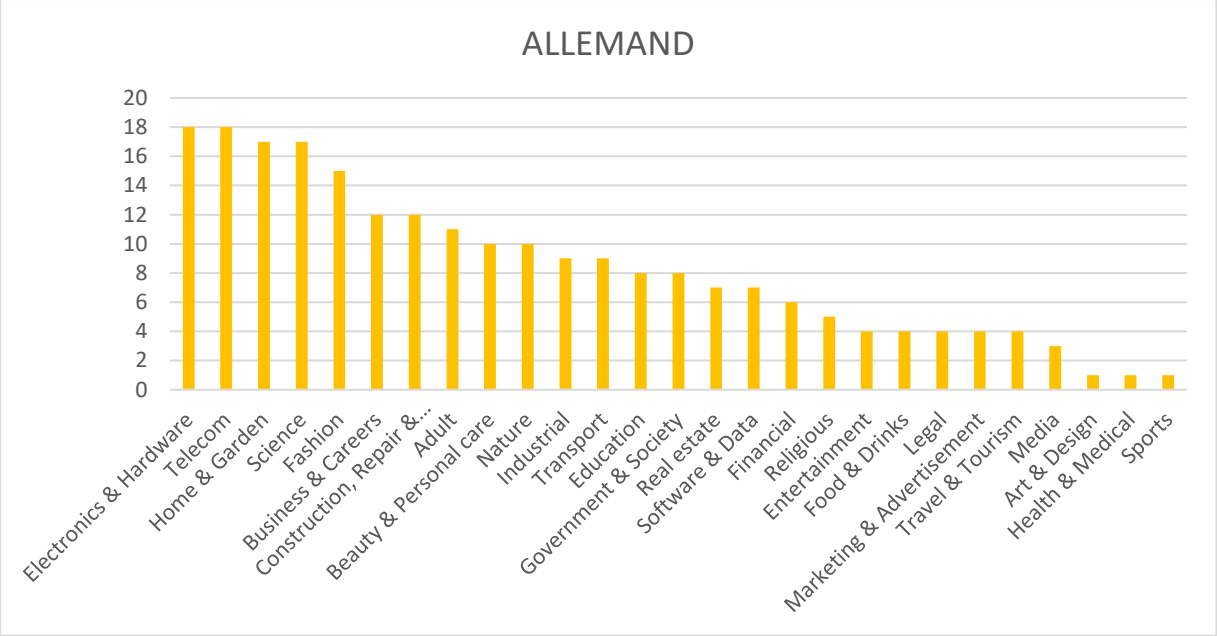
Industrial	
M.POND.	**9,14%**
LANG.	%
Chinese	**22,08%**
Turkish	8,97%
Spanish	8,66%
Japanese	6,25%
Polish	6,03%
Russian	5,69%
Italian	4,27%
Portuguese	4,07%
German	2,65%
Finnish	2,58%
Korean	2,58%
French	1,83%
Hindi	1,71%
English	1,69%
Swedish	1,25%
Norwegian	0,96%
Dutch	0,95%
Danish	0,91%
Hebrew	0,76%

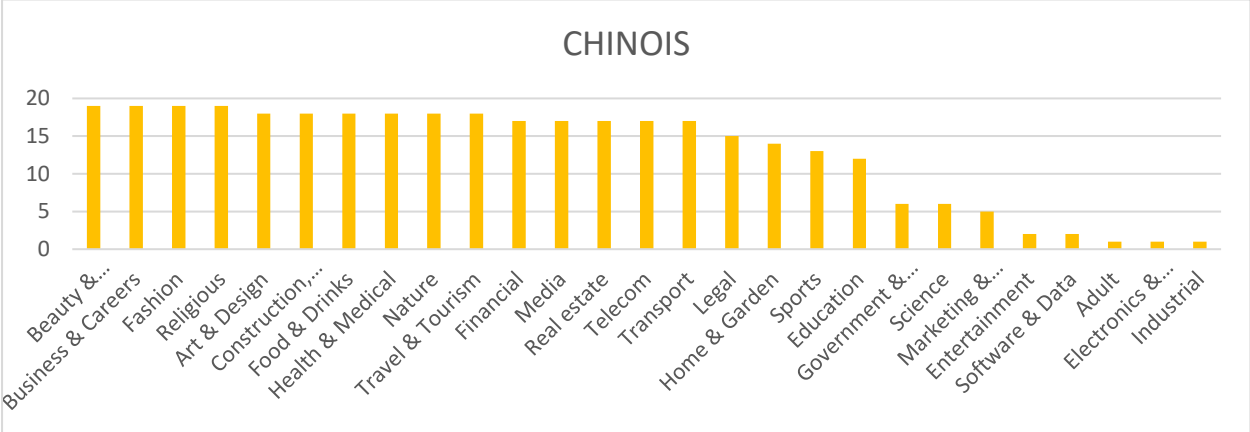
Le chinous écrase les suivants dans le secteur industriel dans des proportions impressionnantes.

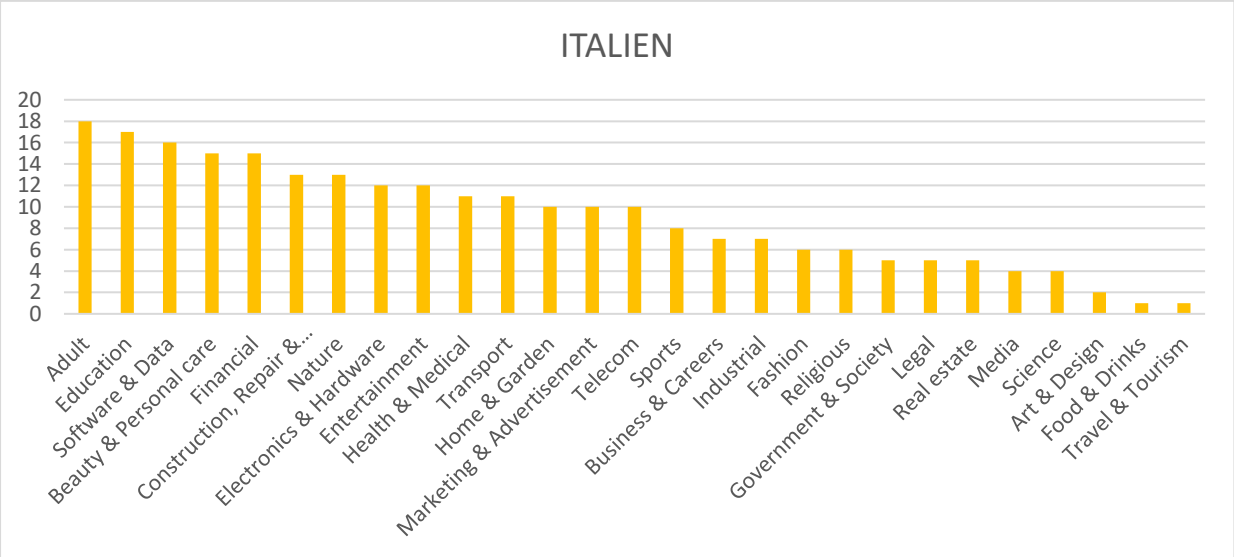
L'analyse ci-dessous de la signature du chinois montre une absence d'équilibre avec cinq thèmes forts et le reste des thèmes dans la partie basse. Mais le pourcentage de site chinois dans la base de données nous parait très en dessous de la réalité et la question du biais possible de sélection reste ouverte.

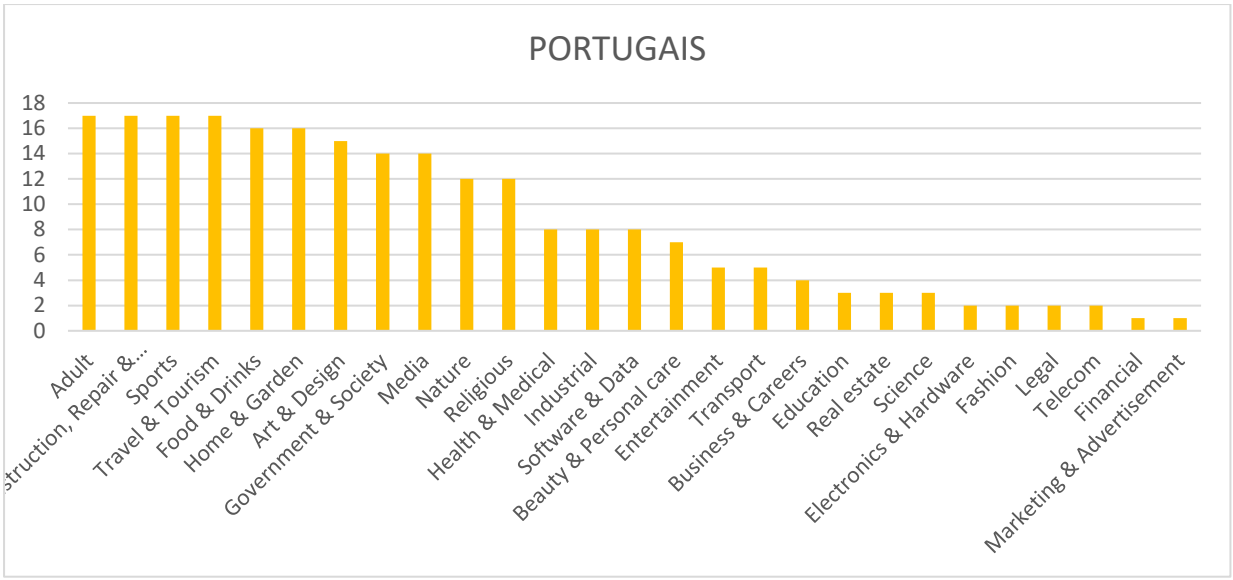
6. 2 Signature thématique de quelques langues

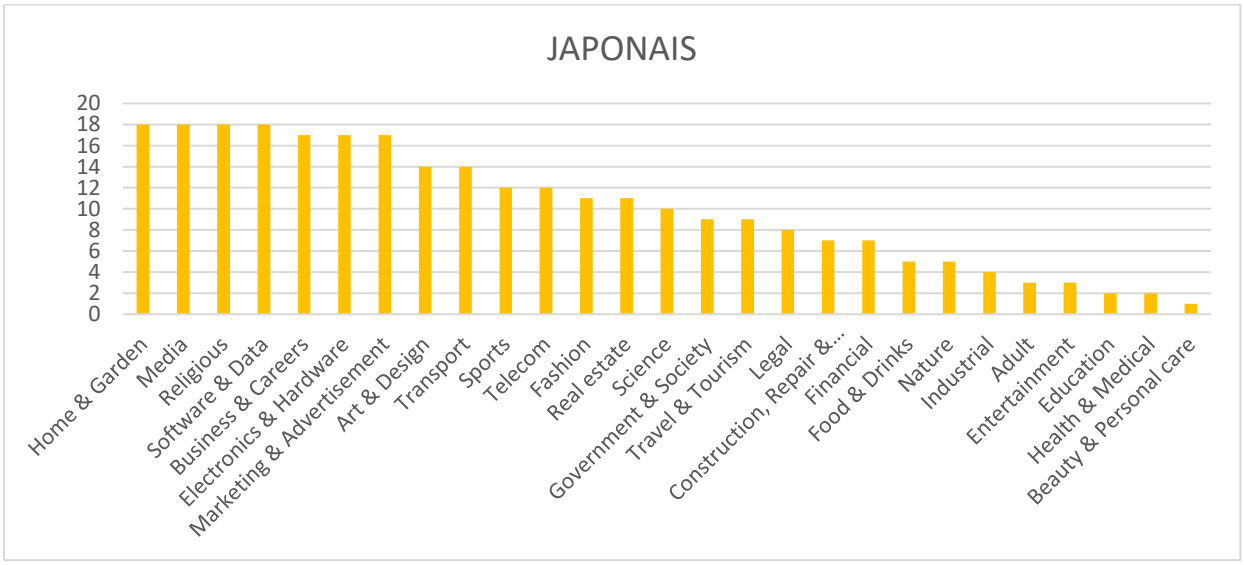


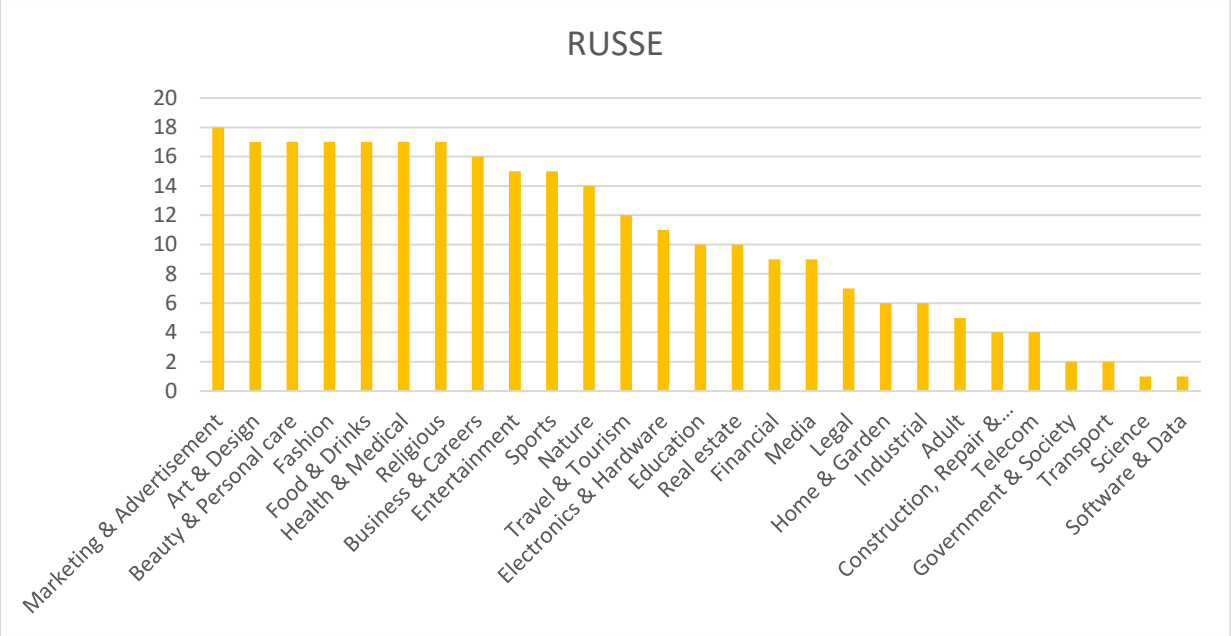

ALLEMAND
20
18
16
14
12
10
8
6
4
2
0
Electronics & Hardware
Telecom
Home & Garden
Science
Fashion
Business & Careers
Construction, Repair &...
Adult
Beauty & Personal care
Nature
Industrial
Transport
Education
Government & Society
Real estate
Software & Data
Financial
Religious
Entertainment
Food & Drinks
Legal
Marketing & Advertisement
Travel & Tourism
Media
Art & Design
Health & Medical
Sports


CHINOIS
20
15
10
5
0
Beauty &...
Business & Careers
Fashion
Religious
Art & Design
Construction,...
Food & Drinks
Health & Medical
Nature
Travel & Tourism
Financial
Media
Real estate
Telecom
Transport
Legal
Home & Garden
Sports
Education
Government &...
Science
Marketing &...
Entertainment
Software & Data
Adult
Electronics &...
Industrial


ITALIEN
20
18
16
14
12
10
8
6
4
2
0
Adult
Education
Software & Data
Beauty & Personal care
Financial
Construction, Repair &...
Nature
Electronics & Hardware
Entertainment
Health & Medical
Transport
Home & Garden
Marketing & Advertisement
Telecom
Sports
Business & Careers
Industrial
Fashion
Religious
Government & Society
Legal
Real estate
Media
Science
Art & Design
Food & Drinks
Travel & Tourism


PORTUGAIS
18
16
14
12
10
8
6
4
2
0
Adult
Construction, Repair &...
Sports
Travel & Tourism
Food & Drinks
Home & Garden
Art & Design
Government & Society
Media
Nature
Religious
Health & Medical
Industrial
Software & Data
Beauty & Personal care
Entertainment
Transport
Business & Careers
Education
Real estate
Science
Electronics & Hardware
Fashion
Legal
Telecom
Financial
Marketing & Advertisement


JAPONAIS
20
18
16
14
12
10
8
6
4
2
0
Home & Garden
Media
Religious
Software & Data
Business & Careers
Electronics & Hardware
Marketing & Advertisement
Art & Design
Transport
Sports
Telecom
Fashion
Real estate
Science
Government & Society
Travel & Tourism
Legal
Construction, Repair &...
Financial
Food & Drinks
Nature
Industrial
Adult
Entertainment
Education
Health & Medical
Beauty & Personal care



6. 3 Matrice des distances thématiques entre langues

	CH	DA	DU	EN	FI	FR	GE	HE	HI	IT	JP	KR	NO	PO	PT	RU	SP	SW	TK
Chinese		7,5	8,7	8,5	8,5	9,0	8,9	9,5	5,5	8,9	7,4	7,4	7,5	9,4	7,8	6,0	8,4	8,2	6,1
Danish	7,5		4,1	7,3	4,9	5,4	6,4	6,6	8,2	6,9	6,3	6,1	5,4	6,6	7,4	7,9	6,0	4,1	6,4
Dutch	8,7	4,1		6,4	4,8	4,7	5,2	6,3	8,0	5,2	6,9	6,2	5,5	6,8	8,1	6,3	5,1	4,4	7,4
English	8,5	7,3	6,4		7,1	4,9	6,1	4,9	9,6	5,5	7,6	6,1	6,4	7,4	5,6	7,7	4,1	7,0	7,9
Finnish	8,5	4,9	4,8	7,1		5,6	4,9	7,8	8,4	6,7	5,9	7,4	5,0	5,4	7,9	6,7	6,2	4,0	6,4
French	9,0	5,4	4,7	4,9	5,6		6,0	7,2	10,2	3,6	6,2	5,9	5,8	6,7	8,0	7,2	5,1	6,0	7,6
German	8,9	6,4	5,2	6,1	4,9	6,0		6,4	8,9	5,2	6,0	6,8	6,6	7,1	7,3	8,2	5,4	6,0	8,5
Hebrew	9,5	6,6	6,3	4,9	7,8	7,2	6,4		9,8	6,7	7,5	6,2	7,3	5,8	5,9	7,3	5,1	7,4	6,3
Hindi	5,5	8,2	8,0	9,6	8,4	10,2	8,9	9,8		9,3	7,6	7,5	7,9	9,5	8,8	7,6	10,1	7,7	7,6
Italian	8,9	6,9	5,2	5,5	6,7	3,6	5,2	6,7	9,3		7,5	4,6	6,3	5,9	7,1	7,0	4,4	6,3	7,3
Japanese	7,4	6,3	6,9	7,6	5,9	6,2	6,0	7,5	7,6	7,5		7,7	6,9	5,9	7,4	6,7	6,5	6,1	5,9
Korean	7,4	6,1	6,2	6,1	7,4	5,9	6,8	6,2	7,5	4,6	7,7		5,9	6,6	6,5	7,8	5,3	6,7	6,3
Norwegian	7,5	5,4	5,5	6,4	5,0	5,8	6,6	7,3	7,9	6,3	6,9	5,9		5,9	7,4	6,9	6,6	4,6	6,7
Polish	9,4	6,6	6,8	7,4	5,4	6,7	7,1	5,8	9,5	5,9	5,9	6,6	5,9		5,9	5,8	5,0	6,1	4,7
Portuguese	7,8	7,4	8,1	5,6	7,9	8,0	7,3	5,9	8,8	7,1	7,4	6,5	7,4	5,9		7,2	4,7	7,7	6,4
Russian	6,0	7,9	6,3	7,7	6,7	7,2	8,2	7,3	7,6	7,0	6,7	7,8	6,9	5,8	7,2		6,5	6,1	4,3
Spanish	8,4	6,0	5,1	4,1	6,2	5,1	5,4	5,1	10,1	4,4	6,5	5,3	6,6	5,0	4,7	6,5		5,7	6,1
Swedish	8,2	4,1	4,4	7,0	4,0	6,0	6,0	7,4	7,7	6,3	6,1	6,7	4,6	6,1	7,7	6,1	5,7		6,2
Turkish	6,1	6,4	7,4	7,9	6,4	7,6	8,5	6,3	7,6	7,3	5,9	6,3	6,7	4,7	6,4	4,3	6,1	6,2	