Profils graphémiques et gabarits abstraits du lexique Ummite (1966-1969) R Galli Novembre 2026

Introduction

Dans l'article « ana50 Langage Ummite : bibliographie et analyses », je concluais que l'abandon de l'hypothèse idéophonémique ouvrait la voie à une étude plus descriptive du corpus Ummite, fondée sur des méthodes quantitatives. Il s'agit désormais de caractériser les formes lexicales observées à partir de propriétés mesurables, avant d'en examiner les éventuelles implications.

Les lexiques des langues présentent des régularités formelles qui peuvent être étudiées statistiquement : fréquence des lettres, distribution des voyelles et des consonnes, récurrence de certaines configurations et diversité des formes. La linguistique de corpus et la linguistique computationnelle proposent différents outils pour décrire ces régularités et comparer des ensembles lexicaux. Toutefois, les résultats obtenus dépendent de la représentation choisie : une analyse fondée sur l'écriture ne renseigne pas directement sur la prononciation, pas plus qu'une distribution statistique ne suffit, à elle seule, à expliquer les mécanismes qui l'ont produite.

Le lexique Ummite présente plusieurs particularités graphiques qui le distinguent au premier abord des lexiques européens étudiés, notamment une proportion élevée de voyelles et une utilisation relativement fréquente des graphèmes W, X et Y. Il convient donc de dépasser cette impression initiale pour examiner quantitativement l'organisation des formes lexicales. L'objectif n'est pas de déterminer d'emblée si ces particularités correspondent à une langue naturelle ou construite, mais de décrire la structure du lexique et de la comparer à celle d'autres corpus au moyen d'une procédure commune.

Les lettres sont ensuite réparties en deux catégories : V pour les graphèmes A, E, I, O et U, et C pour les autres graphèmes retenus dans le corpus. Cette classification est une convention opératoire appliquée aux formes écrites ; elle ne présuppose ni l'inventaire des phonèmes de l'Ummite ni leur réalisation phonétique. Les séquences obtenues sont donc des ***profils graphémiques C/V*** **[1] [2]** et non des transcriptions phonologiques ou des structures syllabiques établies.

Chaque lemme est représenté par son profil C/V, ce qui permet de décrire la diversité des configurations observées et leur fréquence dans le corpus. Ensuite, chaque groupe contigu de consonnes ou de voyelles est ramené à un seul symbole. Cette réduction produit des ***gabarits abstraits*** : par exemple, les profils VVCVV et VVVCCV peuvent être ramenés respectivement à VCV et VCV si leurs groupes contigus sont comprimés selon la règle adoptée. Cela permet ainsi de comparer les formes en faisant abstraction du nombre de graphèmes présents à l'intérieur de chaque groupe, tout en conservant l'ordre des groupes C et V. Cette opération ne mesure ni la longueur phonétique des sons ni le nombre de syllabes des mots.

Les distributions de ces gabarits sont alors caractérisées par leur nombre, leurs fréquences, leur concentration et leur ***entropie de Shannon.*** Ces indicateurs permettent de décrire dans quelle mesure les formes lexicales se répartissent entre plusieurs configurations ou se concentrent sur un nombre restreint de gabarits. L'application de la même procédure à plusieurs autres lexiques fournit un cadre de comparaison quantitatif. Les différences observées doivent toutefois être interprétées en tenant compte de la taille, de la composition et du traitement de chaque corpus. Les différents corpus retenus, au-delà du lexique Ummite retenu, sont basés sur des lexiques de quelques centaines de lemmes en français, en espagnol, en turc, en hawaïen, en esperanto et en langage ro.

Les questions directrices de cette étude sont donc les suivantes :

Quelle organisation statistique des profils graphémiques C/V caractérise le lexique Ummite, et dans quelle mesure cette organisation ressemble-t-elle à celle des lexiques de comparaison étudiés ? Quelles sont les particularités les plus évidentes du corpus Ummite ?

Une précision importante : l'objectif n'est pas de démontrer que le corpus lexical Ummite provient ou non d'une langue naturelle ou artificielle, mais d'évaluer dans quelle mesure ses propriétés statistiques présentent une organisation comparable à celle observée dans les langues naturelles ou artificielles.

Remarque : Cette étude quantitative nécessite l'emploi d'un langage informatique et le développement de différentes fonctions spécialisées. Ces fonctions sont écrites dans le langage Mathematica 15 et sont portées en Annexe 1. Je procéderai de même pour les études suivantes.

[1] Larousse, *Dictionnaire de français*, entrée « graphème ».

[2] Dryer, Matthew S. et Martin Haspelmath (dir.), *WALS Online*, chapitre 12, « Syllable Structure ».

1. Constitution du corpus Ummite

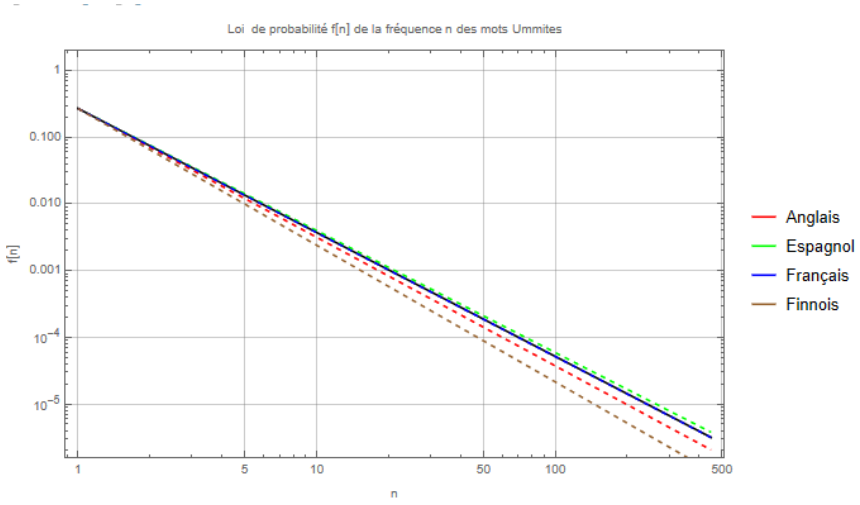
Le corpus de départ utilisé dans cette étude est constitué d'un lexique Ummite élaboré à partir des documents disponibles ; on se basera uniquement sur les vocables trouvés lors de la période 1966-1969, période qui constitue mon sujet d'étude. Le corpus global se trouve sur Ummo-sciences .org et s'intitule « BD VOC.xls ».

A partir de cet ensemble filtré sur la période considérée, on obtient « BD VOC 66-69.xlsx », comprenant plus de 5000 mots, ce qui permet déjà de faire une première analyse en utilisant la loi de Zipf **[3]**. En effet, comme les lettres Ummites sont écrites principalement en espagnol, il serait logique de retrouver la typologie de cette langue dans les lettres car les mots Ummites qui apparaissent dans les lettres sont systématiquement traduits dès les premières lettres et renvoient à un mot ou une explication en espagnol. D'ailleurs, la traduction d'un mot apparait systématiquement lorsqu'il est utilisé la première fois, puis ensuite quelques fois, puis remplace carrément le mot espagnol tellement il apparait maintenant connu et évident par le lecteur attentif.

Comme le montre la graphique suivant, le corpus lexical considéré suit en effet une loi de Zipf **[3]**.

[3] Corral, Á., Boleda, G., & Ferrer-i-Cancho, R. (2015). Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PloS one*, *10*(7), e0129031.

Le lexique « BD VOC 1966-1969 » se répartit selon la loi de Zipf (en noir) qui se confond avec des langues naturelles comme l'espagnol. Comme annoncé, cela est parfaitement logique, car si nous avions pris pour cette étude la traduction de ces mots Ummites en mots espagnols correspondants, on serait arrivés au même résultat



Pour bien comprendre de quoi il s'agit, j'ai réuni les mots les plus souvent rencontrés dans une phrase construite pour l'occasion, et par ordre de leur fréquence rencontrée (entre parenthèses) dans les lettres. Un Ummologue reconnait immédiatement de quoi on parle et sur quels thèmes porte la narration.

« Nous venons de la planète UMMO (449 apparitions). Tout comme vous, nous sommes des OEMMI (162), créés par WOA (156), tous disséminés dans l'infini du WAAM (131). Nos UEWA (102) nous permettent grâce à la théorie des IBOZOO UU (72), de voyager entre les étoiles (la nôtre se nomme IUMMA (45) en des UIW (67) très courts, aidés en cela par nos XANMOO (60) toujours plus miniaturisés. Signé : UMMOAELEWE (59). »

Ainsi même avec un échantillon modeste de 5000 mots, la loi de Zipf se vérifie, indiquant que le lexique restreint sur les années 1966-1969 est statistiquement parfaitement valable. Le lexique obtenu est plus éloigné de l'anglais que d'une langue romane, et encore plus de la typologie d'une langue nordique comme le finnois ainsi que le montre le graphique. Dans la suite, nous ferons souvent référence à l'espagnol.

La fréquence du vocabulaire Ummite employé durant ces 4 premières années révèle également la préoccupation principale de la narration : *la place de l'homme dans l'univers, l'importance de la science et de la technologie qui permettent aux hommes disséminés dans les étoiles de se retrouver dans l'infini du cosmos.*

Il s'agit ensuite de travailler cet échantillon de plus de 5000 mots pour homogénéiser les graphies, corriger certaines incohérences de transcription et supprimer les entrées manifestement redondantes, afin d'obtenir un ensemble de formes lexicales directement exploitable pour une analyse statistique. Ce dernier corpus a besoin d'un nettoyage automatique : des blancs apparaissent en fin ou début de mots, des mots composés sont supprimés automatiquement, ainsi que des mots comprenant des nombres. J'ai supprimé à la main quelques noms propres.

Cette première étape de nettoyage automatique conduit à un corpus de **952** formes lexicales distinctes. Ce corpus correspond donc au matériau documentaire tel qu'il est observé après les opérations élémentaires de normalisation graphique. Il constitue le corpus de référence décrivant fidèlement les formes présentes dans les documents des années madrilènes, de 1966 à 1969.

2. Profil général du corpus Ummite

La première étape consiste à caractériser le profil général du lexique Ummite avant d'étudier l'organisation interne des mots. À ce stade, il ne s'agit pas d'interpréter les résultats mais uniquement de décrire les principales propriétés statistiques du corpus.

2.1 Inventaire et répartition des graphèmes

Le lexique utilise un alphabet relativement restreint composé de vingt-et-un graphèmes :

A, B, C, D, E, G, I, K, L, M, N, O, R, S, T, U, V, W, X, Y et Z.

L'alphabet est donc relativement réduit. Plusieurs lettres de l'alphabet latin sont absentes, notamment **F, H, J, P** et **Q**. Cette première observation montre que le lexique repose sur un système graphique volontairement limité, comparable en taille à celui de nombreuses langues naturelles.

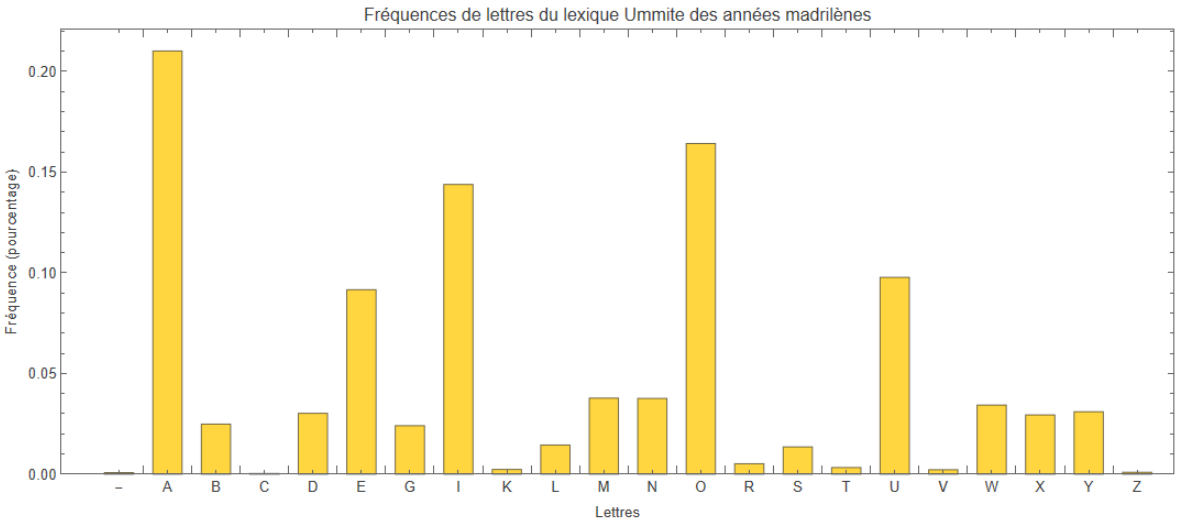
Les graphèmes ne sont pas répartis uniformément. Comme dans les langues naturelles, quelques lettres dominent très largement le corpus tandis que d'autres demeurent peu fréquentes.

Les voyelles **A**, **O** et **I** représentent les graphèmes les plus abondants du lexique. Elles sont suivies par **U** puis **E**, tandis que les consonnes les plus fréquentes sont **M**, **N**, **L**, **R** et **T**.

À l'inverse, les graphèmes **C**, **K**, **V** et **Z** apparaissent rarement.

Les consonnes **W, X et Y**, en revanche, occupent une place beaucoup plus importante que dans la plupart des langues européennes, où elles sont généralement rares.

On peut illustrer graphiquement ce résultat comme suit.



Cette première observation indique déjà que le lexique ne présente pas une distribution uniforme des graphèmes. Comme dans toute langue naturelle, certains graphèmes sont fortement privilégiés alors que d'autres restent exceptionnels.

2.2 Le rapport voyelles / consonnes

Le résultat le plus remarquable concerne la répartition entre voyelles et consonnes. Dans le corpus initial de 952 formes, les voyelles représentent environ **70.8 %** de l'ensemble des caractères. Cela signifie que la très forte proportion de voyelles est une propriété fondamentale du lexique.

La distribution interne des voyelles est également très déséquilibrée. La voyelle **A** représente à elle seule près d'un cinquième des caractères, suivie de **O** puis de **I**, tandis que **U** et **E** occupent des positions plus modestes mais restent néanmoins très fréquentes. Ensemble, ces cinq voyelles constituent la majeure partie des caractères du corpus.

2.3 Les graphèmes W, X et Y : une autre singularité du corpus

Une seconde particularité apparaît lorsque l'on examine les consonnes.

Les graphèmes **W**, **X** et **Y** occupent une place sensiblement plus importante que dans la plupart des langues utilisant l'alphabet latin.

Dans les langues romanes, ces lettres demeurent généralement marginales ou limitées à des emprunts. En anglais, seule la lettre **Y** possède une fréquence relativement élevée, tandis que **W** et **X** restent peu représentées.

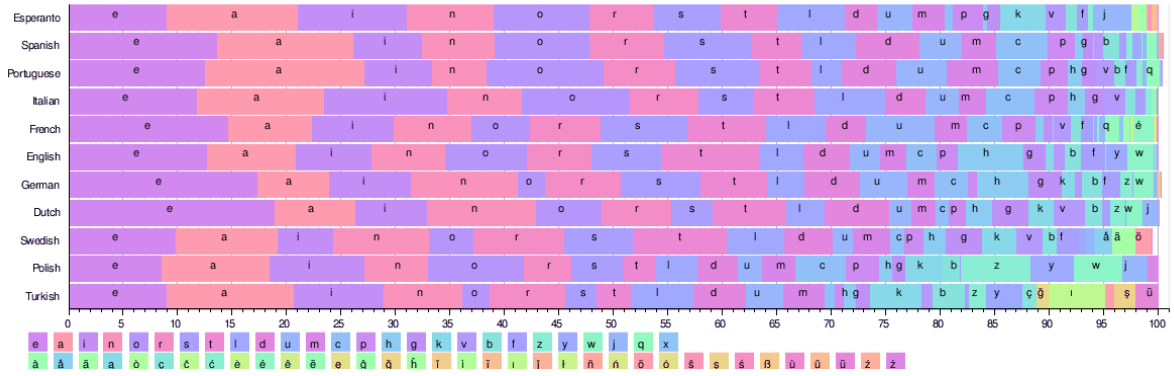
Le lexique Ummite présente une situation différente. Les trois graphèmes apparaissent régulièrement dans les lemmes et participent pleinement à la construction du vocabulaire.

Cette observation est d'autant plus remarquable que l'alphabet utilisé ne comporte que vingt-et-un graphèmes. La fréquence élevée de **W**, **X** et **Y** traduit donc une préférence graphique propre au système et non l'effet d'un alphabet plus étendu.

À ce stade de l'étude, cette particularité constitue un constat statistique. Son éventuelle signification linguistique sera discutée à la lumière des comparaisons avec d'autres langues naturelles et artificielles.

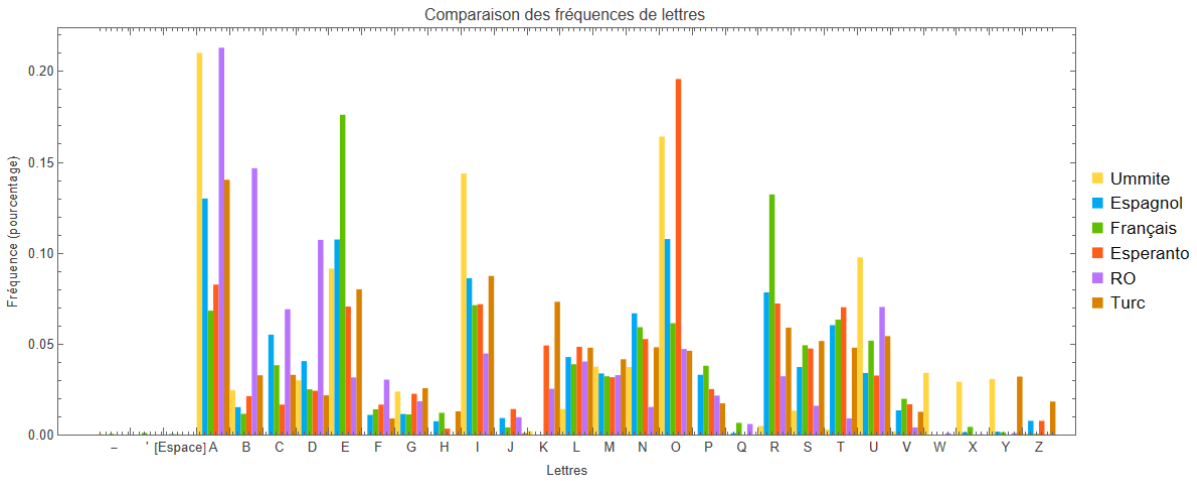
2.4 Un profil atypique en comparaison avec d'autres langues

Afin de situer ce résultat statistique, il est utile de le comparer à la distribution de quelques langues naturelles (Wikipedia /Letter Frequency)



Dans les principales langues indo-européennes, la proportion de voyelles à l'écrit est généralement comprise entre 40 et 50 %. Les langues romanes, comme l'espagnol ou le portugais, présentent des valeurs légèrement supérieures à celles de l'anglais, mais les consonnes y restent majoritaires. Certaines langues austronésiennes, notamment le hawaïen, sont réputées pour leur caractère très vocalique ; néanmoins, leur proportion de voyelles demeure inférieure à celle observée dans le lexique Ummite comme nous le verrons dans le paragraphe suivant.

Voici un graphique illustrant mes résultats comparant le lexique Ummite et différents autres lexiques que j'ai constitués. *Il subsiste quelques espaces et tirets dont la fréquence est non significative.*



On retrouve bien entendu le lexique Ummite en jaune ; on observe effectivement de manière très nette la particularité de ce lexique pour ce qui concerne la fréquence des lettres W, X et Y.

Concernant les taux de voyelles, j'ai observé les taux suivants sur mes propres lexiques : 47 % pour l'espagnol et 43 % pour le français, ce qui est dans la fourchette de la référence ci-dessus. Pour l'hawaïen, j'ai observé 56 % comme prévu pour un langage très vocalique. Et j'ai obtenu 45% pour l'espéranto, 41 % pour le langage ro et enfin 41 % pour le turc.

Ainsi, on peut déjà noter que le lexique Ummite se distingue par une très forte abondance de voyelles inobservée aussi bien dans des langues construites que naturelles.

2.4 Premières synthèses

Cette première analyse met en évidence plusieurs caractéristiques générales :

- Le lexique repose sur un alphabet réduit dont la distribution des graphèmes est fortement hiérarchisée, propriété commune aux langues naturelles.
- On observe une proportion exceptionnellement élevée des voyelles **A, E, I, O, U.**
- On observe également une fréquence inhabituelle des graphèmes W, X et Y, qui occupent une place beaucoup plus importante que dans différentes langues naturelles ou construites.

Ces observations montrent que le lexique Ummite possède une signature statistique originale, suffisamment marquée pour justifier une analyse plus détaillée de son organisation interne.

3. Organisation interne CV du lexique Ummite

3.1 Objectifs de l'analyse et profils graphémiques

Le chapitre précédent a mis en évidence plusieurs caractéristiques générales du lexique Ummite. L'alphabet employé est relativement réduit, la distribution des graphèmes est fortement hiérarchisée, et le corpus se distingue notamment par une proportion exceptionnellement élevée de voyelles ainsi que par une fréquence inhabituelle des graphèmes **W**, **X** et **Y**. Ces observations décrivent la composition du lexique, mais elles ne renseignent pas encore sur son organisation interne.

L'analyse suivante vise à déterminer dans quelle mesure les profils graphémiques du lexique peuvent être regroupés en un nombre limité de gabarits récurrents, puis à comparer la distribution de ces gabarits à celles observées dans plusieurs lexiques de référence.

L'analyse est limitée classiquement à la structure consonantique et vocalique des mots, indépendamment de leur signification ou de leur éventuelle prononciation. Chaque graphème est ainsi remplacé par son appartenance à l'une des deux classes suivantes :

- **V** : voyelle (A, E, I, O, U) ;
- **C** : consonne (tous les autres graphèmes)

Chaque lemme est alors représenté par une succession de symboles **V** et **C**, appelée dans la suite de cette étude **profil graphémique**.

Quelques exemples :

Lemme	**Profil graphémique**
OYAGAA	VCVCVV
BAIXEE	CVVCVV
UMMO	VCCV

Cette représentation conserve uniquement l'alternance entre voyelles et consonnes, sans tenir compte de l'identité des graphèmes. Elle permet ainsi de mettre en évidence les régularités structurelles du lexique indépendamment de son contenu lexical.

L'application de cette transformation aux **952 lemmes** du corpus conduit à l'identification de **277 profils graphémiques distincts**, que l'on désigne par gabarits de surface.

À première vue, ce résultat peut donner l'impression d'une organisation relativement complexe. En effet, près d'un tiers des mots possèdent un profil qui leur est propre, ce qui pourrait laisser penser que les formes lexicales sont distribuées de manière peu contrainte.

Une inspection plus attentive montre cependant que cette diversité est largement superficielle. De nombreux profils ne diffèrent que par la longueur des groupes successifs de voyelles ou de consonnes, alors que leur organisation générale demeure identique.

Ainsi, les profils {VCVV, VVCVV, VVVCVV, VCVVV} présentent des successions différentes de voyelles, mais reposent tous sur la même alternance fondamentale entre un groupe vocalique, une consonne et un nouveau groupe vocalique.

De même, {CVV, CVVV, CVVVV} ne diffèrent que par l'allongement du groupe vocalique final.

Ces premières observations suggèrent que la diversité apparente du corpus provient essentiellement de la longueur variable des groupes vocaliques et consonantiques.

Cette hypothèse conduit naturellement à rechercher un niveau de description plus abstrait, dans lequel seules les alternances entre groupes vocaliques et groupes consonantiques sont conservées. Si cette réduction fait effectivement apparaître un petit nombre de structures récurrentes, elle

constituera un premier indice d'une organisation comparable à celle observée dans les langues naturelles.

3.2 Réduction en gabarits abstraits

Quand les linguistes veulent comparer la structure de langues très différentes (par exemple le français, le turc et l'hawaïen), ils ne peuvent pas se baser sur les mots bruts, ni sur les gabarits de surface précédemment évoqués.

Ils convertissent alors les lexiques en profils graphémiques puis en gabarits abstraits, sorte de squelettes consonantiques et vocaliques de la langue ou du lexique. C'est grâce à cette procédure que l'on pourra ensuite comparer entre elles différentes langues en introduisant certains indicateurs comme par exemple l'entropie de Shannon.

Les gabarits abstraits sont obtenus comme suit : chaque séquence continue de voyelles est remplacée par un unique symbole **V**, tandis que chaque séquence continue de consonnes est remplacée par un unique symbole **C**. Cette transformation conserve l'architecture générale du mot, mais élimine les variations dues à la longueur des groupes.

Par exemple :

Profil graphémique	**Gabarit abstrait**
VCVV	VCV
VVCVV	VCV
VVVCVV	VCV
CVVV	CV
CVVVV	CV
VVCCVVCVV	VCVCV

Les profils graphémiques précédemment distincts apparaissent alors comme des variantes d'une même structure sous-jacente.

L'effet de cette réduction est ici particulièrement remarquable. Les **277 profils graphémiques** observés dans le corpus se ramènent à seulement **17 gabarits abstraits**.

La diversité apparente du lexique diminue ainsi d'un facteur supérieur à seize, ce qui fait apparaître un nombre plus restreint de gabarits abstraits récurrents que ne le laissait supposer l'analyse des profils de surface.

L'examen de la fréquence de ces gabarits montre que leur distribution est très fortement concentrée. Les huit gabarits les plus fréquents couvrent à eux seuls **92 % de l'ensemble du corpus**, tandis que les cinq premiers représentent déjà une très large majorité des lemmes (78%).

Cette forte concentration constitue un premier indice d'une organisation contrainte. Les mots du lexique ne sont manifestement pas construits à partir d'une grande variété d'architectures indépendantes, mais reposent au contraire sur un nombre limité de schémas récurrents.

Une seconde observation apparaît lorsqu'on examine la nature de ces gabarits. Par exemple, les dix structures suivantes ne sont pas indépendantes les unes des autres et appartiennent presque toutes à deux familles très simples.

$$"VCV" \to 260, "VCVCV" \to 203, "CVCV" \to 120, "CV" \to 94, "VCVCVCV" \to 68, "CVCVCV" \to 57, "V" \to 45, "VC" \to 27, "VCVC" \to 17, "VCVCVCVCV" \to 17$$

L'une débute par un groupe vocalique et se termine également par un groupe vocalique :

(VC)^n V qui produit : V, VCV, VCVCV, VCVCVCV, VCVCVCVCV

L'autre correspond à un groupe consonantique :

(CV)^n qui engendre successivement les structures : CVCV, CV, CVCVCV

Ainsi, ces dix gabarits les plus fréquents du corpus Ummite ne constituent pas dix structures indépendantes, mais dérivent essentiellement de deux types de régularités extrêmement simples.

La richesse apparente des profils graphémiques ne traduit pas une grande liberté combinatoire ; elle résulte principalement de l'allongement variable des groupes vocaliques, qui multiplie les profils de surface sans modifier leur architecture fondamentale.

Le lexique présente ainsi une organisation hiérarchique à deux niveaux. Au premier niveau, les profils graphémiques décrivent fidèlement les successions de voyelles et de consonnes et rendent compte de la diversité observable des mots. Au second niveau, les gabarits abstraits révèlent un nombre très restreint de structures fondamentales, communes à la quasi-totalité du corpus.

Cette distinction entre formes de surface et gabarits sous-jacents constitue le résultat principal de l'analyse. Elle montre que le lexique n'est pas organisé comme un ensemble arbitraire de combinaisons graphiques, mais qu'il repose sur une distribution concentrée sur un nombre limité de gabarits, dans laquelle un petit nombre de modèles structurels suffit à rendre compte de la presque totalité des lemmes étudiés.

Il convient toutefois de souligner que cette observation ne constitue pas, à elle seule, une propriété spécifique du lexique Ummite. De nombreuses langues limitent également la construction de leurs mots à un nombre restreint de gabarits comme nous allons le voir.

La question n'est donc plus de savoir si le corpus présente une organisation, mais de déterminer si cette organisation est comparable à celle observée dans différents lexiques. C'est l'objet de la section suivante.

3.3 Comparaison avec différents lexiques

L'analyse précédente montre que la réduction ramène les 277 profils observés à 17 gabarits abstraits. Cela permet ainsi d'étudier la répartition des formes entre un ensemble plus restreint de configurations. Une telle organisation pourrait constituer une propriété remarquable du corpus. Toutefois, elle ne prend réellement son sens que si elle est comparée à d'autres corpus lexicaux issus de langues naturelles ou artificielles, l'ensemble étant étudié selon une méthodologie identique.

Différents lexiques figurant en Annexe 2 (Ummite, Espagnol, Français, Esperanto, Ro, Turc, Hawaïen) ont été soumis exactement au même traitement selon le déroulé suivant :

1. Conversion de chaque lemme en profil graphémique V/C et réduction des groupes successifs de voyelles et de consonnes à un unique symbole ;
2. Dénombrement des gabarits abstraits ;
3. Calcul de leur fréquence cumulée : par exemple, quelle proportion du lexique est couverte par les 8 premiers gabarits abstraits;
4. Estimation de l'entropie de Shannon et de l'indice de contrainte.

Pourquoi tous ces traitements et notamment le point 4 ? Le simple nombre de gabarits abstraits ne renseigne pas complètement sur la structure globale du lexique. Deux langues peuvent posséder le même nombre de structures fondamentales tout en les distribuant de manière très différente.

Pour cette raison, une estimation de l'entropie de Shannon a été réalisée. L'entropie, calculée sur la distribution des gabarits abstraits et non sur les graphèmes eux-mêmes, mesure la diversité effective d'une distribution : elle est faible lorsque quelques gabarits dominent largement le corpus et augmente lorsque les fréquences tendent à s'équilibrer.

Le calcul de l'entropie de Shannon est complété par différents autres paramètres associés, le tout aboutissant à ce qu'on appelle l'indice de contrainte ou de redondance.

Le tableau suivant explique ce que sont ces différents paramètres et ce qu'ils montrent.

Terme	**Signification & Définition**	**Ce que cela montre**
Entropie de Shannon (H)	Mesure de la quantité moyenne dite d'incertitude ou d'information (en bits) $H = -\sum_i [p_i \log_2(p_i)]$	Plus H est bas, plus le lexique est concentré sur un petit nombre de gabarits.
Entropie maximale (Hmax)	Entropie théorique d'un système où tous les gabarits auraient la même probabilité : $Hmax = \log_2(N)$	C'est la capacité combinatoire théorique maximale offerte par le nombre de gabarits identifiés.
Entropie relative (H / Hmax)	Rapport entre l'incertitude observée et l'incertitude maximale possible (exprimé en %).	C'est la capacité combinatoire relative offerte par le nombre de gabarits identifiés.
Indice de contrainte / Redondance (R)	Taux de structuration : $R = 1 – (H)/Hmax$	Mesure de l'écart relatif entre l'entropie observée et l'entropie maximale.

Ceci étant rappelé, voici le tableau récapitulatif obtenu calculant pour chaque lexique les différents paramètres précédemment décrits :

Langue	Gabarits abstraits (réduits)	Couverture des 8 principaux gabarits	Entropie de Shannon H (gabarits abstraits)	Entropie Hmax	Taux de contrainte
Ummite	17	91.8	3.08	4.09	25.
Espagnol	22	85.6	3.27	4.46	27.
Français	23	93.	3.08	4.52	32.
Espéranto	14	98.8	2.30	3.81	40.
RO	8	100.	1.48	3.00	50.6
Turc	19	89.3	3.15	4.25	26.
Hawaien	25	89.	3.12	4.64	33.

Précisons également qu'un nombre plus élevé de gabarits abstraits indique une plus grande diversité de configurations observées, mais pas nécessairement une langue plus complexe.

La couverture des huit premiers gabarits donne le pourcentage des lemmes représentés par les huit configurations les plus fréquentes. Plus ce pourcentage est élevé, plus une fraction importante du lexique est concentrée sur un petit ensemble de gabarits. Ce choix a été fait en raison du nombre très limité de gabarits abstraits du langage ro.

La couverture des huit premiers gabarits et le taux de contrainte sont en fait complémentaires : la première se concentre sur les huit formes dominantes, tandis que le second prend en compte l'ensemble de la distribution.

3.4 Commentaires sur le tableau comparatif précédent

Le tableau précédent permet de comparer les lexiques selon trois dimensions complémentaires : le nombre de gabarits C/V abstraits observés, la part du lexique couverte par les huit gabarits les plus fréquents et la distribution des occurrences entre les gabarits, mesurée par l'entropie de Shannon et le taux de contrainte.

L'Ummite : une diversité de gabarits associée à une concentration modérée

Le lexique Ummite présente 17 gabarits abstraits. Les huit plus fréquents couvrent 91,8 % du corpus, ce qui révèle une forte réutilisation d'un ensemble relativement restreint de configurations. Toutefois, cette concentration n'empêche pas une diversité appréciable des formes : l'entropie atteint 3,08 bits, pour un maximum de 4,09 bits, soit un taux de contrainte de 25 %.

Ce résultat place l'Ummite dans le groupe des corpus où la distribution des gabarits est relativement étalée. Son entropie est identique à celle du français et proche de celles de l'espagnol, du turc et de l'hawaïen. Son taux de contrainte est même légèrement inférieur à ceux de ces quatre corpus.

Le point intéressant est donc la coexistence de deux propriétés : une forte couverture par les gabarits dominants et une distribution qui conserve une diversité statistique notable. Les formes lexicales ne se répartissent pas uniformément entre les 17 gabarits, mais elles ne sont pas non plus concentrées aussi fortement que dans les corpus d'espéranto et de RO.

Ces observations caractérisent la distribution graphémique du corpus. Elles ne permettent pas, à elles seules, d'affirmer que les gabarits correspondent à des règles de construction des mots.

L'espagnol : une distribution plus diversifiée

Le corpus espagnol comporte 22 gabarits abstraits, soit davantage que l'Ummite. Les huit gabarits les plus fréquents couvrent 85,6 % des lemmes, la couverture la plus faible parmi les corpus présentés. L'entropie atteint 3,27 bits, la valeur la plus élevée de l'ensemble, tandis que le taux de contrainte s'établit à 27 %.

Cette combinaison indique une distribution relativement étalée entre un nombre important de configurations. Le lexique espagnol présente ici une plus grande diversité de gabarits observés que le lexique Ummite, et les formes sont moins concentrées dans les huit configurations principales.

Cette différence doit être interprétée avec prudence : le nombre de gabarits dépend notamment de la composition du corpus et des formes lexicales retenues. Il montre en revanche que, selon la représentation adoptée, son lexique présente une distribution plus diversifiée que celle de l'Ummite.

Le français : une diversité élevée, mais une concentration plus marquée

Le français présente 23 gabarits, soit le deuxième nombre le plus élevé après l'hawaïen. Les huit premiers couvrent 93 % du lexique. Son entropie est de 3,08 bits, identique à celle de l'Ummite, mais son taux de contrainte atteint 32 %, contre 25 % pour l'Ummite.

Cette apparente coexistence entre un nombre élevé de gabarits, une couverture importante des huit premiers et une entropie modérée illustre l'intérêt de combiner les indicateurs. Le nombre de configurations renseigne sur la diversité des formes possibles observées ; il ne dit rien, à lui seul, de la façon dont les occurrences se répartissent entre elles.

Le français possède donc, dans ce corpus, davantage de gabarits que l'Ummite, mais sa distribution est plus concentrée au regard du taux de contrainte. L'égalité des entropies ne signifie pas que les deux distributions sont identiques : plusieurs distributions distinctes peuvent produire la même entropie.

L'espéranto : une forte concentration sur un petit nombre de gabarits

L'espéranto ne présente que 14 gabarits abstraits, le deuxième nombre le plus faible du tableau. Les huit premiers couvrent 98,8 % du lexique. Son entropie est de 2,30 bits et son taux de contrainte atteint 40 %.

Il s'agit d'une distribution plus concentrée que celles des quatre langues naturelles comparées et de l'Ummite, ceci sur les corpus étudiés. La quasi-totalité des lemmes est représentée par huit configurations, et les fréquences sont suffisamment inégales pour que l'entropie soit sensiblement inférieure à son maximum.

L'espéranto étant une langue construite, on pourrait être tenté d'expliquer ce résultat par la régularité de sa conception. Mais ce tableau ne permet pas d'établir ce lien causal. Il faudrait notamment comparer d'autres langues construites et contrôler la composition des corpus. La conclusion directement soutenue par les chiffres est plus limitée : le corpus d'espéranto étudié présente une distribution des gabarits plus concentrée que les autres corpus du tableau, selon les indicateurs retenus.

RO : la distribution la plus concentrée

Le corpus RO se distingue nettement des autres avec seulement huit gabarits abstraits. Les huit premiers couvrent donc 100 % du lexique. Son entropie est de 1,48 bit, pour un maximum de 3 bits, et son taux de contrainte atteint 50,6 %, valeur la plus élevée du tableau.

La combinaison d'un petit nombre de gabarits et d'une entropie relativement faible indique une distribution fortement concentrée. Le lexique RO mobilise un répertoire de configurations restreint et les occurrences ne se répartissent pas uniformément entre ces configurations.

Il serait toutefois excessif d'en conclure que RO est une langue plus simple ou plus artificielle que les autres. Ces indicateurs ne mesurent ni la complexité grammaticale ni la richesse sémantique. Ils caractérisent exclusivement la distribution des gabarits graphémiques dans le corpus retenu.

Le turc : une distribution relativement étalée

Le corpus turc comporte 19 gabarits abstraits. Les huit premiers couvrent 89,3 % du lexique, l'entropie atteint 3,15 bits et le taux de contrainte est de 26 %.

Le turc se situe ainsi près de l'espagnol et de l'Ummite en matière de concentration statistique. Sa distribution est plus étalée que celles du français, de l'espéranto et de RO selon le taux de contrainte, et son entropie est proche de celle de l'espagnol.

Le turc possède par ailleurs une structure morphologique agglutinante, avec des suffixes qui peuvent s'enchaîner. Cependant, on ne peut pas attribuer les résultats observés à cette caractéristique sans une analyse spécifique. Les données présentées ne permettent pas de savoir en effet quelle part de la distribution résulte des propriétés morphologiques du turc.

L'hawaïen : beaucoup de gabarits, mais une concentration intermédiaire

L'hawaïen présente 25 gabarits abstraits, le nombre le plus élevé du tableau. Les huit principaux couvrent 89 % du lexique, tandis que l'entropie atteint 3,12 bits et le taux de contrainte 33 %.

Le corpus hawaïen combine donc une diversité importante de configurations observées et une concentration non négligeable des occurrences sur les huit premières. Son taux de contrainte est supérieur à ceux de l'Ummite, de l'espagnol et du turc, mais inférieur à ceux de l'espéranto et de RO.

Rappelons que les gabarits ne sont pas des structures syllabiques, mais des représentations abstraites des séquences graphémiques de mots entiers. Cette distinction est essentielle pour éviter de confondre les résultats du tableau avec une analyse phonologique de la langue.

3.5 Synthèse comparative

Trois groupes se dessinent dans les résultats.

- Distribution relativement étalée : l'Ummite, l'espagnol et le turc présentent les taux de contrainte les plus faibles, entre 25 et 27 %. Leurs distributions répartissent les occurrences entre plusieurs gabarits, même si les nombres de gabarits diffèrent.
- Situation intermédiaire : le français et l'hawaïen affichent des taux de contrainte de 32 et 33 %. Ils possèdent un nombre relativement élevé de gabarits, mais leurs distributions sont plus concentrées que celles du premier groupe.
- Distribution fortement concentrée : l'espéranto et RO se distinguent avec des taux de contrainte de 40 et 50,6 %, associés à un nombre plus faible de gabarits et à une couverture presque totale par les huit premiers.

Ce classement reste descriptif. Il ne constitue pas une typologie générale des langues naturelles et construites, puisque les corpus sont peu nombreux.

3.6 En guise de conclusion du chapitre 3 : le lexique Ummite est-il atypique ?

Petite précaution préliminaire : les corpus de comparaison constituent des échantillons lexicaux de référence et non des inventaires exhaustifs construits selon un protocole de sélection identique. Les différences observées doivent donc être interprétées comme des différences entre les corpus étudiés, et non comme des estimations définitives des propriétés de chaque langue.

Ceci étant précisé, les résultats obtenus indiquent que le lexique Ummite présente une organisation graphémique C/V caractérisée par un répertoire de 17 gabarits abstraits, dont les huit plus fréquents couvrent 91,8 % des lemmes. Malgré cette forte couverture, la distribution conserve une entropie de 3,08 bits et un taux de contrainte de 25 %, ce qui traduit une concentration modérée par rapport aux autres corpus étudiés.

Selon ces indicateurs, l'Ummite se situe dans le même ordre de grandeur que l'espagnol et le turc, et présente une distribution plus étalée que celles observées pour l'espéranto et RO. Il se distingue notamment de l'espéranto par un nombre plus élevé de gabarits, une couverture moins forte par les huit premiers gabarits et une entropie supérieure. Ces différences constituent des observations quantitatives, mais elles ne suffisent pas à établir une opposition générale entre langues naturelles et langues construites.

La conclusion la plus solide est donc que le lexique Ummite étudié n'apparaît pas atypique au regard des seuls indicateurs de distribution graphémique retenus. Sa structure statistique présente des ressemblances avec plusieurs corpus de langues naturelles, sans être identique à aucun d'entre eux.

Cela justifie de poursuivre l’étude linguistique du lexique selon d’autres facettes, notamment sa morphologie.

Conclusion

Cette étude avait pour objectif de caractériser quantitativement les profils graphémiques C/V du lexique Ummite de la période 1966-1969 et de comparer leur distribution à celles observées dans plusieurs lexiques de référence.

La procédure appliquée aux 952 formes retenues conduit à 277 profils graphémiques distincts, regroupés en 17 gabarits abstraits après compression des groupes contigus de voyelles et de consonnes. Les huit gabarits les plus fréquents couvrent 91,8 % du corpus. L'entropie de Shannon atteint 3,08 bits, pour une entropie maximale de 4,09 bits, soit un taux de contrainte d'environ 25 % selon la définition retenue.

Ces indicateurs décrivent une distribution dans laquelle un nombre limité de gabarits représente une grande partie des formes lexicales, tout en conservant une diversité statistique appréciable. Comparé aux corpus de référence étudiés, le lexique Ummite présente des valeurs d'entropie et de contrainte proches de celles obtenues pour plusieurs langues naturelles, tandis que les corpus d'espéranto et de RO présentent des distributions plus concentrées. Ces différences concernent les corpus examinés et ne permettent pas, à elles seules, de caractériser les langues naturelles ou construites dans leur ensemble.

L'analyse des graphèmes met également en évidence deux particularités du corpus Ummite : une proportion élevée de voyelles, estimée à 70,8 % des caractères, et une fréquence notable des graphèmes W, X et Y. Ces observations portent sur les formes écrites et ne permettent pas de déduire directement leur réalisation phonétique ou leur fonction linguistique.

En définitive, cette étude propose une méthode descriptive de comparaison des profils graphémiques et de leurs distributions. Elle ne permet pas d'établir l'origine du corpus, ni de démontrer l'existence de règles phonologiques ou morphologiques sous-jacentes. Elle fournit en revanche un premier ensemble de résultats quantitatifs qui pourra être confronté à d'autres dimensions du lexique.

L’étape suivante consistera à examiner, de manière également quantitative, l’organisation morphologique du même corpus (familles lexicales, mécanismes de formation des lemmes, relations entre formes).

Remarque

Au-delà des résultats obtenus, la succession de fonctions Mathematica portées en Annexe 1 permet de reproduire les résultats de cet article en utilisant les lexiques de l’Annexe 2. Le lecteur pourra aussi utiliser d’autres lexiques de son choix.

Annexe 1 Fonctions spécifiques développées en langage Mathematica pour cette étude

Pour les besoins de cette étude, il est très important d'avoir les données d'entrée, c'est-à-dire chaque lemme, comme suit : sans accents, en caractères majuscules, et de plus sans blancs. On modifiera dans cette étude les mots composés soit en supprimant les blancs soit en les décomposant pour les ajouter au lexique.

Exemples utilisant la fonction resultats décrite plus loin :

resultats[DeleteDuplicates[StringDelete[lexic,Whitespace]]] pour supprimer les blancs

resultats[DeleteDuplicates[Flatten[StringSplit[lexic]]]] pour décomposer les mots composés.

```
(*Calcul de la fréquence des lettres apparaissant dans un lexique  lexic_List*)

FrequenceLettres[lexic_List]:=Module[{dist,tot},

 dist=ReverseSort[Merge[CharacterCounts[lexic],Total]];

 tot=Total[dist];

 N[dist/tot,3] ]  ]


(*Fonction qui remplace les voyelles par V et les consonnes par C. Rappelons que le Y est considéré comme une consonne en espagnol*)

FormeDuMot[mot_String]:=StringReplace[ToUpperCase[mot],{"A"|"E"|"I"|"O"|"U"->"V","B"|"C"|"D"|"F"|"G"|"H"|"J"|"K"|"L"|"M"|"N"|"P"|"Q"|"R"|"S"|"T"|"V"|"W"|"X"|"Z"|"Y"->"C"}];

phono[lexic_List]:=ReverseSort[Counts[Table[FormeDuMot[lexic[[i]]],{i,1,Length[lexic]}]]]


(*compression des groupes contigus de voyelles et de consonnes*)

simplifyPattern[str_]:=StringReplace[str,{Repeated["V",{2,20}]->"V",Repeated["C",{2,20}]->"C"}];

(*Regroupement trié*)

groupedByAbstractShape[lexic_List]:=ReverseSort[Merge[KeyValueMap[simplifyPattern[#1]->#2&,phono[lexic]],Total]];

(* Affichage des résultats *)

resultats[lexic_List]:=Module[{},

 Print["\n--- STRUCTURES ABSTRAITES APRÈS COMPRESSION DES DOUBLONS (V.. -> V) ---"];

 Print["Nombre de patrons initiaux : ",Length[phono[lexic]]];

 Print["Nombre de patrons abstraits réduits : ",Length[groupedByAbstractShape[lexic]]];

 Print["\nLes 10 structures abstraites les plus fréquentes :"];

 Print[Take[groupedByAbstractShape[lexic],10]];]
```

```
(*Calcul de l'entropie de Shannon sur le lexique*)
ShannonLexique[lexic_List]:=Module[{totalTokens,lexmots,probs,shannonEntropy,nTypes,maxEntropy,relativeEntropy,redundancy},
  lexmots=lexic;
 (*1. Calcul des probabilités de chaque patron abstrait*)
 totalTokens=Total[Values[groupedByAbstractShape[lexmots]]];
 probs=Values[groupedByAbstractShape[lexmots]]/totalTokens;
 (*2. Entropie de Shannon (H) en bits*)
 shannonEntropy=N[-Total[probs*Log2[probs]],3];
 (*3. Entropie maximale théorique (H_max)*)
 nTypes=Length[groupedByAbstractShape[lexmots]];
 maxEntropy=N[Log2[nTypes],3];
 (*4. Entropie relative et Taux de Redondance/Contrainte*)
 relativeEntropy=shannonEntropy/maxEntropy;
 redundancy=(1-relativeEntropy)*100;
 {nTypes,shannonEntropy,maxEntropy,relativeEntropy,redundancy}
 ]
(*Taux de concentration des gabarits abstraits sur les p premiers*)
Concentre[gabarit_, p_] := Module[{freqs, totalTokens, cumulPercent},
 freqs = Values[gabarit];
 totalTokens = Total[gabarit];
 cumulPercent = N[Accumulate[freqs]/totalTokens*100];
 Round[cumulPercent[[p]], 0.1]
 ]
```

Exemples de résultats sur un lexique français nommé FR

ffr=FrequenceLettres[FR]

<|"E"->0.176,"R"->0.132,"I"->0.0716,"A"->0.0685,"T"->0.0636,"O"->0.0617,"N"->0.0595,"U"->0.0520,"S"->0.0495,"L"->0.0391,"C"->0.0385,"P"->0.0383,"M"->0.0325,"D"->0.0253,"V"->0.0201,"F"->0.0143,"H"->0.0124,"B"->0.0118,"G"->0.0116,"Q"->0.00688,"X"->0.00468,"J"->0.00440,"Y"->0.00165,"'"->0.00138,"-"->0.000826,"Z"->0.000826," "->0.000551|>

Lookup[ffr, {"A","E","I","O","U"}];Total[%] (*Lookup est une fonction intégrée dans MMA 15*)

0.430

resultats[FR]

"--- STRUCTURES ABSTRAITES APRÈS COMPRESSION DES DOUBLONS (V.. -> V) ---"

Nombre de patrons initiaux : 160

Nombre de patrons abstraits réduits : 23

Les structures abstraites les plus fréquentes :

<|"CVCVC"->154,"CVCV"->125,"CVC"->102,"CVCVCVC"->51,"CVCVCV"->43,"VCVCVC"->35,"VCVC"->34,"VCV"->14,"VCVCV"->12,"CV"->6|>

groupedByAbstractShape[FR]

<|"CVCVC"->154,"CVCV"->125,"CVC"->102,"CVCVCVC"->51,"CVCVCV"->43,"VCVCVC"->35,"VCVC"->34,"VCV"->14,"VCVCV"->12,"CV"->6,"CVCVCVCV"->6,"VCVCVCVC"->5,"VC"->2,"C'VCVC"->2,"VC-CVCV"->1,"CVC-VCV"->1,"CVCV'VC"->1,"CVCV CV"->1,"VCVC'CV"->1,"CV-CVC"->1,"C'VC"->1,"VCVCVCV"->1,"CVCVCVCVC"->1|>

Concentre[groupedByAbstractShape[FR],8]

93.

shfr=ShannonLexique[FR]

{23,3.08,4.52,0.68,32.}

Les graphiques et les tableaux du corps de l'article sont basés sur les données numériques issues des fonctions spécifiquement développées pour cette étude et qui sont appliquées à différents lexiques.

Ces résultats ne prétendent pas être représentatifs d'un lexique bien plus volumineux : ils reposent sur des lexiques limités de quelques centaines de lemmes comme montré en Annexe 2.

Annexe 2 Présentation des différents lexiques choisis pour cette étude.

Français

fr={"être","avoir","vous","aller","faire","pour","dire","pouvoir","vouloir","mais","nous","dans","elle","savoir","bien","voir","plus","avec","devoir","tout","venir","rien","comme","suivre","parler","prendre","cette","votre","quand","alors","chose","croire","aimer","falloir","très","quoi","passer","penser","aussi","jamais","attendre","pourquoi","trouver","laisser","arriver","donner","regarder","encore","appeler","est-ce que","homme","partir","petit","toujours","jour","femme","temps","maintenant","notre","deux","mettre","rester","sans","arrêter","vraiment","connaître","merci","tuer","comment","mourir","demander","même","peut-être","dieu","fois","père","comprendre","seul","quelque","sortir","trop","chez","fille","monde","vrai","autre","besoin","accord","monsieur","enfant","entendre","voilà","chercher","heure","aider","mère","déjà","beau","essayer","quel","juste","mieux","grand","quelqu'un","beaucoup","revenir","donc","plaire","maison","gens","après","nuit","personne","soir","bonjour","jouer","finir","peur","parce que","perdre","maman","sentir","ouais","rentrer","avant","problème","argent","quelle","vivre","rendre","tenir","main","cela","vite","oublier","depuis","fils","travailler","moins","tête","coup","écouter","raison","manger","mort","amour","entrer","devenir","hein","commencer","moment","dernier","voiture","demain","elles","payer","tirer","ouvrir","oeil","fait","changer","question","tomber","assez","foutre","excuser","affaire","dormir","combien","frère","travail","idée","famille","truc","trois","tant","premier","tous","occuper","entre","marcher","chance","aujourd'hui","envoyer","histoire","tard","apprendre","minute","boire","garder","quelques","type","porte","montrer","asseoir","porter","année","sous","souvenir","prêt","contre","prier","pendant","mois","servir","madame","putain","écrire","part","sang","place","espérer","désoler","jeune","retrouver","ville","terre","gagner","semaine","acheter","longtemps","gars","chambre","côté","leur","vieux","rappeler","lire","seulement","voici","salut","monter","désolé","quitter","police","suite","matin","emmener","toucher","continuer","gros","enfin","mari","là-bas","revoir","importer","papa","puis","train","manquer","raconter","ensemble","mauvais","film","répondre","garçon","corps","celui","autres","heureux","loin","sauver","chaque","retourner","leurs","rencontrer","coeur","voler","fermer","valoir","descendre","docteur","suffire","façon","nouveau","important","sembler","compter","vers","joli","point","hier","chien","dont","guerre","genre","merde","meilleur","marier","arme","cause","endroit","ordre","poser","reste","pied","envie","tiens","inquiéter","bouger","plutôt","apporter","pardon","photo","plein","devant","décider","ainsi","droit","aucune","vendre","école","cher","chef","tourner","cacher","boulot","pays","ceux","possible","expliquer","battre","peine","livre","agir","imaginer","tour","adorer","vérité","recevoir","gentil","jeter","pleurer","bébé","partie","nouvelle","amener","instant","parent","service","plaisir","promettre","mentir","soeur","bientôt","tellement","utiliser","lieu","coucher","presque","dehors","préférer","content","pauvre","offrir","verre","réveiller","aucun","aide","d'abord","journée","préparer","numéro","permettre","ramener","enlever","chéri","fort","lâcher","choisir","musique","conduire","faute","calmer","mariage","bureau","route","chanter","disparaître","lever","présenter","accepter","sinon","long","café","propre","confiance","cinq","bonsoir","compte","téléphone","casser","prochain","frapper","facile","attention","rêve","copain","malade","lettre","ignorer","fête","couper","esprit","super","taire","seigneur","flic","parfois","capitaine","autant","force","pièce","quatre","difficile","bras","tromper","surtout","ressembler","jurer","plan","prison","sale","courir","état","remettre","carte","paix","noir","exactement","drôle","refuser","président","cours","grave","terminer","ciel","partout","patron","amuser","celle","visage","intéresser","reconnaître","rire","médecin","rapport","pardonner","avis","embrasser","retour","simple","danser","exister","différent","souvent","lumière","génial","libre","près","dollar","sûrement","pareil","d'autres","hôpital","ceci","réussir","bizarre","voix","protéger","honneur","équipe","prévenir","cheval","ailleurs","maître","avion","bout","habiter","faim","ensuite","l'un","froid","normal","déranger","reprendre","oncle","prix","retard","détruire","cadeau","pousser","face","gueule","chemin","vivant","général","bateau","million","impossible","seconde","découvrir","erreur","soleil","voyage","sauter","rêver","détester","clair","faux","paraître","balle","empêcher","maintenir","cheveu","papier","sujet","supposer","tranquille","épouser","blanc","table","toute","agent","approcher","sens","craindre","message","crier","salle","inviter","effet","réfléchir","espèce","arranger","naître","allô","passé","bois","propos","camp","sorte","hôtel","début","souffrir","jambe","remercier","choix","sécurité","baiser","avocat","attraper","client","courant","dépêcher","peuple","dame","dessus","abandonner","vérifier","journal","sérieux","brûler","fond","gosse","fric","situation","sauf","accident","doute","scène","soldat","amoureux","assurer","preuve","humain","silence","télé","victime",

"complètement","pute","calme","derrière","garde","meurtre","groupe","crime","traiter","doucement","blesse
r","colonel","certain","secret","parole","triste","honte","rouge","dégager","soirée","dangereux"};

Pour le français, il faut supprimer les diacritiques et passer en majuscules avant l'aspect algorithmique.

FR=ToUpperCase@*RemoveDiacritics/@fr

Espagnol

esp={"ANO","DIA","PERSONA","HOMBRE","MUJER","CASA","TIEMPO","VIDA","MANO","PARTE","MUNDO","PA
IS","CIUDAD","FAMILIA","AGUA","TRABAJO","DINERO","PROBLEMA","FORMA","MOMENTO","CASO","LUGAR",
"COSA","NINO","NINA","PADRE","MADRE","HIJO","HIJA","AMIGO","AMIGA","GRUPO","HISTORIA","PUEBLO","
CALLE","ESCUELA","LIBRO","PALABRA","PREGUNTA","RESPUESTA","VERDAD","IDEA","RAZON","DERECHO","LE
Y","GUERRA","PAZ","FUERZA","PODER","GOBIERNO","PRESIDENTE","MINISTRO","REY","REINA","SOCIEDAD","E
MPRESA","NEGOCIO","MERCADO","PRECIO","VALOR","NUMERO","NOMBRE","EDAD","FIN","PRINCIPIO","INICI
O","CAMBIO","NIVEL","EJEMPLO","IMAGEN","VOZ","SONIDO","MUSICA","LUZ","COLOR","NOCHE","MANANA"
,"TARDE","SEMANA","MES","HORA","MINUTO","SEGUNDO","CUERPO","CABEZA","CARA","OJO","BOCA","NARI
Z","OREJA","PELO","BRAZO","PIERNA","PIE","DEDO","CORAZON","SANGRE","SALUD","ENFERMEDAD","MEDIC
O","HOSPITAL","COMIDA","PAN","ARROZ","CARNE","PESCADO","FRUTA","VERDURA","LECHE","QUESO","HUE
VO","AZUCAR","SAL","CAFE","TE","VINO","CERVEZA","MESA","SILLA","PUERTA","VENTANA","SUELO","TECHO",
"PARED","HABITACION","COCINA","BANO","CAMA","ROPA","ZAPATO","CAMISA","PANTALON","BOLSILLO","B
OLSA","PAPEL","CARTA","PERIODICO","REVISTA","TELEFONO","ORDENADOR","PANTALLA","RADIO","TELEVISIO
N","COCHE","TREN","AVION","BARCO","BICICLETA","CAMINO","CARRETERA","PUENTE","RIO","MAR","OCEAN
O","LAGO","MONTANA","BOSQUE","ARBOL","FLOR","HOJA","RAIZ","CIELO","SOL","LUNA","ESTRELLA","LLUVIA
","NIEVE","VIENTO","FUEGO","TIERRA","ARENA","PIEDRA","ANIMAL","PERRO","GATO","CABALLO","VACA","CE
RDO","OVEJA","CABRA","PAJARO","PEZ","INSECTO","ABEJA","MARIPOSA","SERPIENTE","UNIVERSIDAD","PROF
ESOR","ESTUDIANTE","CLASE","EXAMEN","CIENCIA","ARTE","CULTURA","LENGUA","IDIOMA","LITERATURA","P
OESIA","NOVELA","EDIFICIO","PISO","BARRIO","ZONA","REGION","ESTADO","PROVINCIA","FRONTERA","CAPIT
AL","CONTINENTE","ISLA","PLAYA","CAMPO","GRANJA","JARDIN","PARQUE","PLAZA","TIENDA","RESTAURANT
E","BAR","HOTEL","BANCO","OFICINA","FABRICA","TALLER","ESTACION","AEROPUERTO","PUERTO","DIRECCIO
N","ENTRADA","SALIDA","LLAVE","ESCALERA","ASCENSOR","DORMITORIO","SALON","COMEDOR","ARMARIO",
"MANTA","ALMOHADA","LAMPARA","VASO","PLATO","TAZA","BOTELLA","CUCHILLO","TENEDOR","CUCHARA",
"OLLA","SARTEN","ALIMENTO","DESAYUNO","ALMUERZO","CENA","POLLO","MANZANA","NARANJA","LIMON"
,"TOMATE","PATATA","CEBOLLA","AJO","ACEITE","HARINA","PASTEL","CHOCOLATE","HELADO","PROFESION","
EMPLEO","CARRERA","JEFE","EMPLEADO","COMPANERO","CLIENTE","SERVICIO","PRODUCTO","VENTA","COM
PRA","CONTRATO","SALARIO","COSTE","CUENTA","IMPUESTO","RECURSO","EDUCACION","CONOCIMIENTO","
APRENDIZAJE","MEMORIA","ATENCION","INFORMACION","DATO","TEXTO","PAGINA","CAPITULO","ARTICULO"
,"FRASE","LETRA","MENSAJE","NOTICIA","CUENTO","RELATO","DOCUMENTO","FIRMA","FOTO","PELICULA","J
UEGO","DEPORTE","EQUIPO","PARTIDO","JUGADOR","PELOTA","FUTBOL","CANCION","INSTRUMENTO","GUIT
ARRA","PIANO","BAILE","TEATRO","CINE","PINTURA","DIBUJO","OBRA","ARTISTA","ACTOR","ACTRIZ","ESCRIT
OR","POETA","CREACION","TECNOLOGIA","MAQUINA","PROGRAMA","SISTEMA","RED","INTERNET","APLICACI
ON","APARATO","ENERGIA","ELECTRICIDAD","GAS","METAL","MADERA","PLASTICO","MATERIAL","MATERIA","
ESPACIO","AIRE","ATMOSFERA","CLIMA","TEMPERATURA","PRIMAVERA","VERANO","OTONO","INVIERNO","M
UERTE","NACIMIENTO","INFANCIA","JUVENTUD","VEJEZ","SUENO","MIEDO","AMOR","ALEGRIA","TRISTEZA","
FELICIDAD","EMOCION","SENTIMIENTO","ESPERANZA","DUDA","SORPRESA","LIBERTAD","RESPONSABILIDAD",
"RELACION","SITUACION","CONDICION","MANERA","CAUSA","EFECTO","RESULTADO","PROCESO","DESARROLL
O","MOVIMIENTO","OCASION","OPORTUNIDAD","POSIBILIDAD","NECESIDAD","DECISION","ELECCION","ACUE
RDO","OPINION","PROPUESTA","PLAN","PROYECTO","OBJETIVO","META","SOLUCION","ERROR","DIFICULTAD",
"RIESGO","EXITO","FRACASO","EXPERIENCIA","PRACTICA","ACTIVIDAD","ESFUERZO","AYUDA","APOYO","CONT
ROL","MEDIDA","CANTIDAD","CALIDAD","TAMANO","PESO","DISTANCIA","VELOCIDAD","POSICION","PUNTO",
"LINEA","LADO","CENTRO","EXTREMO","TIPO","ESPECIE","GENERO","CATEGORIA","FIGURA","ESTRUCTURA","

BASE","ELEMENTO","CONJUNTO","UNIDAD","MITAD","TOTAL","CIUDADANO","GENTE","COMUNIDAD","POBL
ACION","NACION","POLITICA","VOTO","AUTORIDAD","POLICIA","EJERCITO","SEGURIDAD","CRIMEN","JUSTICIA
","JUEZ","TRIBUNAL","RELIGION","IGLESIA","TRADICION","COSTUMBRE","FIESTA","CELEBRACION","DIA
FESTIVO","PAREJA","MATRIMONIO","HERMANO","HERMANA","ABUELO","ABUELA","TIO","TIA","PRIMO","PRI
MA","VECINO","ADULTO","JOVEN","PERSONA
MAYOR","PIEL","HUESO","MUSCULO","CEREBRO","ROSTRO","CUELLO","HOMBRO","PECHO","ESTOMAGO","ES
PALDA","DOLOR","MEDICINA","TRATAMIENTO","OPERACION","CUIDADO","PLANTA","SEMILLA","TERRENO","N
ATURALEZA","MEDIO
AMBIENTE","RAMA","FRUTO","FUENTE","COSTA","VALLE","ROCA","HUMO","NUBE","TORMENTA","CALOR","F
RIO","SOMBRA","PLANETA","UNIVERSO","OBJETO","HERRAMIENTA","PIEZA","COMPONENTE","MODELO","VE
RSION","DISENO","SUPERFICIE","BORDE","INTERIOR","EXTERIOR","RUTA","VIAJE","DESTINO","LLEGADA","VISIT
A","TURISTA","VIAJERO","TRANSPORTE","VEHICULO","MOTOR","RUEDA","CONDUCTOR","TRAFICO","AVENIDA
","TUNEL","AUTOBUS","COMUNICACION","LENGUAJE","EXPRESION","CONVERSACION","DISCUSION","REUNIO
N","LLAMADA","CORREO","SENAL","INFORME","ARCHIVO","LISTA","REGISTRO","TITULO","CURSO","LECCION",
"ALUMNO","PROFESORA","NOTA","PENSAMIENTO","HABILIDAD","CAPACIDAD","TALENTO","METODO","TECNI
CA","INVESTIGACION","ESTUDIO","DESCUBRIMIENTO","PRUEBA","ECONOMIA","MONEDA","COMERCIO","TRA
BAJADOR","PRODUCCION","INDUSTRIA","ORGANIZACION","COMPANIA","LIDER","MIEMBRO","CONTACTO","B
EBE","ANCIANO","PACIENTE","MENTE","CONFIANZA","INTENCION","MENTIRA","REALIDAD","FINAL","FUTURO
","PASADO","PRESENTE","EPOCA","VEZ","RECUERDO","IMAGINACION","CREATIVIDAD","VOLUNTAD","DESEO",
"ODIO","AMISTAD","RESPETO","IGUALDAD","DIFERENCIA","DEBER","REGLA","NORMA","PELIGRO","VENTAJA",
"ALTURA","LONGITUD","EJERCICIO","DESCANSO","PASEO","AVENTURA","TERRITORIO","AMBIENTE","VESTIDO"
,"SOMBRERO","RELOJ","FOTOGRAFIA"};

Hawaïen

haw={"WAI","KAI","ʻĀINA","HALE","KANAKA","KĀNE","WAHINE","KEIKI","MAKUAKĀNE","MAKUAHINE","ʻOHA
NA","KUPUNA","MOʻOPUNA","KAIKAINA","KAIKUAHINE","KAIKUAʻANA","HOAALOHA","HOA","MALIHINI","ALIʻ
I","MAKAʻĀINANA","KONOHIKI","KAHUNA","KUMU","KAMA","INOA","MOʻOLELO","ʻŌLELO","MELE","OLI","HU
LA","PULE","MANA","ALOHA","MAHALO","PONO","MĀLAMA","LAʻA","HEIAU","AKUA","AUMĀKUA","MOʻO","K
UPUA","LANI","HONUA","AO","PŌ","LĀ","MAHINA","HŌKŪ","NALU","KAHAKAI","ONE","PŌHAKU","MAUNA","
PUʻU","LUA","AWĀWA","KAHAWAI","PUNAWAI","LOKO","LOKOIʻA","WAILELE","UA","MAKANI","ANUENUE","P
UA","LAU","LĀʻAU","ʻAʻA","ʻULU","NIU","KALO","MAIʻA","MĀLA","LOʻI","ʻAI","MEAʻAI","IʻA","MANŌ","PUAʻA",
"MOA","ʻĪLIO","PŌPOKI","LIO","HOLOHOLONA","MANU","NALO","PULELEHUA","POʻO","MAKA","IHU","WAHA
","NIHO","PEPEIAO","LAUOHO","ʻILI","IWI","LIMA","WĀWAE","MANAMANA","NAʻAU","PUʻUWAI","KOKO","OL
A","MAʻI","LĀʻAU LAPAʻAU","KAUKA","HALE
MAʻI","MOE","MOEMOEĀ","HAUʻOLI","KAUMAHA","MAKAʻU","HUHŪ","ʻIKE","MANAʻO","NAʻAUAO","HAʻAWI
NA","KULA","KUMU
KULA","HAUMĀNA","PALAPALA","PUKE","ʻAOʻAO","PEPA","LEKA","NŪPEPA","KIʻI","KIʻIʻONIʻONI","KANI","LEO"
,"ʻOHE","IPU","PAHU","KAULA","WAʻA","HOE","LAWAIʻA","UPENA","KAPA","MALO","PAʻU","LEI","ʻAHU","MEA
","WAIWAI","WAIHONA","HALE
KŪʻAI","KĀLĀ","KŪʻAI","HANA","ʻOIHANA","LAWELAWE","MĪKINI","KAʻA","ALANUI","ALA","PUKA","PUKA
ANIANI","PĀ","PAPA","MOENA","NOHO","IPUHAO","HANAUNA","MOKUʻĀINA","KULANAKAUHALE","KAUHALE
","MĀLAʻAI","KAIĀULU","LĀHUI","AUPUNI","KANĀWAI","MALUHIA","KULEANA","NOHONA","ʻIKEPILI","HŌʻAIL
ONA","INOA WAHI","WAHI","MANAWA","MAKAHIKI","MOKU","ʻĀINA HĀNAU","WAIHONA
ʻIKE","PILI","HĀLĀWAI","KIPA","HUAKAʻI","MOANA","KAIĀ","KIHI","WAENA","MUA","HOPE","MAKUA","KEIKI
KĀNE","KEIKI WAHINE","PŌKIʻI","HĀNAU","KUPUNA KĀNE","KUPUNA
WAHINE","ʻANAKALA","ʻANAKĒ","HOAHĀNAU","HOA PILI","HOA HANA","HOA NOHO","MAKUA
KĀNE","MAKUA WAHINE","ʻOHANA PILI","HALE NOHO","HALE MOE","HALE ʻAINA","HALE PULE","HALE
HANA","HALE KULA","HALE ALIʻI","PUKA KOMO","PĀ HALE","PAPAHELE","PĀPALE","IPU
WAI","PUNA","PAHI","KĪ","KĪʻAHA","KAʻAAHI","MOKULELE","UWAPO","AWA","KAHUA","WAHO","KALANA","A
HUPUAʻA","WAHI PANA","PALI","ANA","WAIHONA WAI","MOKUPUNI","WAO","WAO
AKUA","KO","ʻUALA","ʻŌHIʻA","HONU","NĒNĒ","KINO","MANAMANA

LIMA","KULI","PO‘OHIWI","UMAUMA","‘ŌPŪ","MĀLAMA KINO","HO‘OMANA","PAHU
HULA","HO‘OKIPA","HELE","HOLO","MAKAU","AHO","KO‘A","PALAOA","MŌ‘Ī","MŌ‘ĪWAHINE","KĀNAKA
MAOLI","KAHU","LIMAHANA","HANA
LIMA","MĀKEKE","KUMUKŪ‘AI","UKŪ","‘AI‘Ē","LOA‘A","HUI","‘AHAHUI","‘AHA","KŪKĀKŪKĀ","NĪNAU","PANE"
,"A‘O","KULANUI","HUA‘ŌLELO","MĀHELE","PĀ‘ANI","PILA","PA‘I
KI‘I","NO‘EAU","KĀMA‘A","‘EKE","PŪ‘OLO","MEA HANA","PĀKAUKAU","MEA
KANU","KOA","KUKUI","‘UHANE","HANU","KAU","WĀ","MĀLAMA
‘ĀINA","KAPU","NOA","LONO","KU","KANALOA","PELE","HI‘IAKA","HĀLAU","KAHUNA LAPA‘AU","KAHUNA
PULE","‘AUMAKUA","AKAMAI","HO‘ONA‘AUAO","HALE WAIHONA PUKE","MO‘OKŪ‘AUHAU","‘ŌLELO
HAWAI‘I","IPU
HEKE","UKULELE","‘OLU‘OLU","LOKAHI","HO‘OMANAWANUI","MAKEMAKE","‘OIA‘I‘O","WAHAHE‘E","HO‘OMA
NA‘O","MANAOLANA","LAPA‘AU","HO‘ŌLA","MĀMĀ","‘ŌPŪ LI‘ILI‘I","HALE
KUKE","KŌKUA","ALAKA‘I","LUNA","KANAWAI","HO‘OKOLOKOLO","‘AE","HŌ‘OLE","KAKAHIAKA","AHIAHI","A‘A
","KOHOLĀ","MELE HULA"};

Pour l’hawaïen, et pour l'analyse CV, il faut remplacer le kahako par une voyelle double.

Quant au okina, il agit comme une consonne un peu spéciale : nous allons le remplacer par un X, sachant qu'il n'y a pas de X dans le présent lexique hawaïen, ce qui permettra de ne pas oublier de consonnes.

HAW=StringReplace[haw,{(*Macrons hawaïens*)"Ā"->"AA","Ē"->"EE","Ī"->"II","Ō"->"OO","Ū"->"UU",(*Okina
*)"‘"->"X"}] ;

Turc Le nettoyage des diacritiques a déjà été fait pour le turc.

turc={"AT","KUS","BALIK","KOPEK","KEDI","AYI","KURT","KOYUN","KAZ","ASLAN","BOCEK","YILAN","TAVUK","E
SEK","ARI","FARE","KUZU","KELEBEK","TAVSAN","FIL","DEVE","KECI","GUVERCIN","MARTI","SINEK","INEK","KA
RINCA","KARTAL","MAYMUN","YENGEC","TILKI","BOGA","DOMUZ","AKREP","LEYLEK","GEYIK","KAPLUMBAGA"
,"KURBAGA","KARGA","OKUZ","ORUMCEK","CEYLAN","SERCE","HOROZ","KAPLAN","SIGIR","ORDEK","BAYKUS",
"HINDI","TIMSAH","MANDA","KUGU","SINCAP","SOLUCAN","YARASA","CEKIRGE","SALYANGOZ","ATMACA","Z
URAFA","AHTAPOT","EL","BAS","YUZ","GOZ","DIL","AYAK","AGIZ","KOL","SAC","KALP","BEYIN","KULAK","PAR
MAK","DIS","SIRT","BURUN","BACAK","DUDAK","OMUZ","ALIN","GOGUS","SINIR","BEL","BOGAZ","BOYUN","K
EMIK","DIZ","DAMAR","KARIN","MIDE","KAS","YANAK","CILT","KAS","TIRNAK","GOBEK","CENE","AKCIGER","BI
LEK","RAHIM","BOBREK","BAGIRSAK","KARACIGER","KALCA","DIRSEK","KIRPIK","TOPUK","KAFATASI","OMURG
A","BASPARMAK","KABURGA","EKMEK","YUMURTA","UN","BAL","PEYNIR","ZEYTIN","PIRINC","YOGURT","SIMI
T","KASAR","PEKMEZ","CACIK","CEKIRDEK","DOMATES","SOGAN","ELMA","PATATES","KIRAZ","MANTAR","BIB
ER","SALATA","PORTAKAL","CEVIZ","INCIR","KABAK","FINDIK","FASULYE","HAVUC","KAYISI","FISTIK","KARPUZ"
,"ERIK","KESTANE","PATLICAN","ARMUT","BADEM","MUZ","CILEK","VISNE","LAHANA","ISPANAK","HIYAR","AY
VA","SEFTALI","KAVUN","ENGINAR","MANDALINA","BARBUNYA","TAVUK","HAMSI","SUCUK","KIYMA","KARIDE
S","BIFTEK","PALAMUT","PASTIRMA","KUSBASI","PIRZOLA","LEVREK","CIPURA","KALAMAR","YAG","SEKER","T
UZ","TEREYAGI","KARABIBER","NANE","SIRKE","KEKIK","TARCIN","KIRMIZIBIBER","HARDAL","KIMYON","PASTA
","DONDURMA","HELVA","KEK","KURABIYE","BAKLAVA","MUHALLEBI","ASURE","SUTLAC","KOMPOSTO","GULL
AC","REVANI","KUNEFE","KESKUL","BARDAK","KASIK","TABAK","BICAK","FINCAN","CATAL","KASE","DONER","C
ORBA","KOFTE","PILAV","BOREK","KISIR","MAKARNA","KEBAP","DOLMA","MANTI","LAHMACUN","HUMUS","T
ANDIR","GUVEC","ERISTE","SU","CAY","KAHVE","SUT","SARAP","BIRA","RAKI","KOLA","AYRAN","SERBET","GAZ
OZ","IHLAMUR","SODA","BOZA","SALEP","CEP","ELBISE","AYAKKABI","KIYAFET","ETEK","DUGME","GOMLEK","
PANTOLON","SAPKA","CEKET","CORAP","KEMER","KAZAK","KOT","SEMSIYE","PALTO","TISORT","KRAVAT","KO
STUM","TERLIK","BOT","UNIFORMA","GECELIK","ELDIVEN","SORT","PIJAMA","HIRKA","MAYO","MONT","BLUZ"
,"YELEK","BERE","ATKI","ESARP","ESOFMAN","KASKET","SUTYEN","BIKINI","KULOT","FERMUAR","YAGMURLUK
","PAPYON","KASKOL","SUVETER","SAAT","CANTA","GOZLUK","YUZUK","MENDIL","CUZDAN","KOLYE","BILEZIK
","AKSESUAR","KUPE","YOL","SOKAK","HEDEF","GEMI","SEFER","YOLCULUK","TRAFIK","HARITA","YOLCU","CA
DDE","SEYAHAT","LIMAN","KAPTAN","SURUCU","PILOT","KALDIRIM","UCUS","TUNEL","HAVAALANI","DEMIRY
OLU","ROTA","PIST","GEZINTI","OTOYOL","BULVAR","GARAJ","HAVALIMANI","SEYYAH","GUZERGAH","OTOGA

R","GAR","OTOBAN","NAVIGASYON","ARAC","ARABA","UCAK","GEMI","OTOBUS","TREN","OTOMOBIL","TAKSI
","KAMYON","HELIKOPTER","BISIKLET","MINIBUS","DOLMUS","METRO","DENIZALTI","BOT","AMBULANS","TRA
MVAY","MOTORSIKLET","KAMYONET","FERIBOT","UZAYGEMISI","KIZ","BABA","ANNE","OGUL","KARDES","ABI"
,"HANIM","KOCA","KARI","ES","HALA","ABLA","DEDE","AMCA","GELIN","TEYZE","OGLAN","DAYI","EVLAT","AN
NEANNE","TORUN","BABAANNE","DAMAT","NINE","BIRADER","YEGEN","ENISTE","YENGE","BACI","KUZEN","EB
EVEYN","KAYNANA","BUYUKANNE","KAYINVALIDE","BUYUKBABA","VALIDE","KAYINPEDER","GORUMCE","PEDE
R","KAYINBIRADER","BALDIZ","BACANAK","ELTI","KAYINBABA","GUVEY","TAS","ALTIN","PETROL","GAZ","CAM"
,"DEMIR","KAGIT","CELIK","KURSUN","KUM","METAL","YAKIT","KOMUR","PLASTIK","GUMUS","KUL","ODUN","
IPEK","AHSAP","KRISTAL","CIMENTO","BAKIR","MERMER","LASTIK","INCI","KIL","SERAMIK","KARTON","TUGLA
","YUN","SAMAN","ELMAS","ASFALT","BRONZ","ALUMINYUM","PORSELEN","KETEN","CILA","KAUCUK","GRANI
T","KOTON","PAZAR","CUMA","CUMARTESI","PAZARTESI","CARSAMBA","PERSEMBE","SALI","GECE","SABAH","
AKSAM","OGLE","GUNDUZ","AKSAMUSTU","IKINDI","GECEYARISI","OCAK","ARALIK","EYLUL","MAYIS","EKIM","
MART","NISAN","KASIM","TEMMUZ","HAZIRAN","AGUSTOS","SUBAT","YAZ","KIS","SONBAHAR","ILKBAHAR","Y
IL","GUN","AY","SAAT","HAFTA","YUZYIL","DAKIKA","MEVSIM","SANIYE","CEYREK","TAKVIM","MILENYUM","SA
LISE","OYUN","TAKIM","MAC","TOP","SPOR","SAHA","GOL","YARIS","HAVUZ","KUPA","FORMA","ATIS","MADA
LYA","ANTRENMAN","SAMPIYONA","STADYUM","MUSABAKA","MARATHON","KORT","KANO","RAKET","FUTBO
L","BASKETBOL","TENIS","KAYAK","VOLEYBOL","BOKS","ATLETIZM","KARATE","BEYZBOL","CIMNASTIK","HENTB
OL","HOKEY","JUDO","ESKRIM","OYUNCU","FUTBOLCU","HAKEM","SAMPIYON","GALIP","KALECI","MAGLUP","
ATLET","BOKSOR","DALGIC","YUZUCU","BASKETBOLCU","KOSUCU","SORFCU","TENISCI","DOVUSCU","VOLEYB
OLCU","CIMNASTIKCI","GOLFCU","YER","DUNYA","SU","CEVRE","TOPRAK","DOGA","ARAZI","ATMOSFER","MA
NZARA","AY","DENIZ","GUNES","BITKI","BAHCE","CICEK","DAG","AGAC","YILDIZ","ORMAN","ADA","TEPE","KIYI
","DALGA","GOK","GOL","GOKYUZU","ZIRVE","NEHIR","UFUK","KUM","OT","SAHIL","GEZEGEN","COL","VADI","
KORFEZ","OVA","YOKUS","VOLKAN","CALI","AKINTI","YARIMADA","KUMSAL","CIMEN","TAHIL","EGIM","PLATO
","SELALE","GUNES","YAGMUR","KAR","BULUT","SEL","RUZGAR","FIRTINA","ATMOSFER","SIS","SIMSEK","TAYF
UN","HORTUM","KASIRGA","TIPI","GOKGURULTUSU","SICAK","SOGUK","BUZ","DON","ILIK","NEMLI","KARLI","
GUNESLI","KURAK","YAGMURLU","AYAZ","DONDURUCU","BULUTLU","RUZGARLI","HAVA","SICAKLIK","IKLIM","
MEVSIM","METEOROLOJI","ASKER","YAZAR","OGRETMEN","DOKTOR","POLIS","MESLEK","MEMUR","SAIR","AV
UKAT","SUBAY","SAVCI","MUHENDIS","SOFOR","HEKIM","MIMAR","IMAM","SURUCU","CIFTCI","POLITIKACI","
HEMSIRE","BAKKAL","BEKCI","COBAN","GARSON","BALIKCI","BERBER","AKADEMISYEN","HIZMETCI","EGITIMCI
","DIPLOMAT","EBE","ASCI","KASAP","TERZI","DENIZCI","CEVIRMEN","KUAFOR","KITAPCI","YAYINCI","TEKNISY
EN","BAHCIVAN","DEMIRCI","BOYACI","KUYUMCU","VETERINER","FIRINCI","MARANGOZ","BARMEN","DISCI","
BALERIN","AYAKKABICI","SAATCI","ITFAIYECI","DOGRAMACI","ASCIBASI","DOKUMACI","MATBAACI","BAHARAT
CI","BEYAZ","KARA","YESIL","KIRMIZI","SIYAH","MAVI","AK","PEMBE","KIZIL","MOR","GRI","KAHVERENGI","SA
RI","LACIVERT","TURUNCU","BORDO","BEJ","TURKUAZ","HAKI","CAMGOBEGI","ZEYTUNI","SIKLAMEN","ET","CI
K","GOR","ANLA","BUL","GEC","YAP","AL","TUT","OL","TANI","SEV","CEK","BIL","VER","YASA","KAC","KARSILA"
,"SOYLE","YAZ","DE","AYIR","KULLAN","DUY","OKU","CALIS","AT","SUR","GIR","ARA","OL","DUR","GETIR","BAK
","DEGIS","AC","KONUS","GEL","DUS","ULAS","GIT","YE","KAL","UY","KES","ISTE","KUR","DUSUN","OYNA","BA
SLA","BEKLE","KAPA","KAZAN","BIRAK","DON","KOY","YURU","GOSTER","IN","SIK","KAT"};

Esperanto

ranto={"BOVLO","PATO","KALDRONO","STUFUJO","FORNO","KUIRFORNO","FRIDUJO","LAVILO","VAKUILO","SA
KO","KOFRO","KESTO","SKATOLO","TUBO","FADENO","PINGLO","TONDILO","BROSO","KOMBILO","TOLUJO","LI
TTUKO","KAPKUSENO","MATRACO","KURTENO","TAPETO","TAPESTRO","VAZO","KANDELABRO","KANDELO","A
RBUSTO","BRANCO","TIGO","RADIKO","FOLIO","SEMO","BURGONO","BOSKO","STEPO","DEZERTO","GANGALO
","MARCO","LAGETO","KAVERNO","KLIFO","VULKANO","SISMO","INUNDO","TEMPESTO","FULMO","TONDRO",
"ROSO","FROSTO","NEBULO","TORNADO","HURAGANO","RADIADO","MEDIO","LEPORO","KUNIKLO","VULPINO
","LUPO","APRO","CERVO","URSETO","SIMIO","ELEFANTO","TIGRO","ZEBRO","GIRAFO","KAMELO","KROKODIL
O","TESTUDO","RANO","PAPILIO","ABELO","VESPO","FORMIKO","ARANEO","MUSO","MISKITO","VERMO","KO
NKO","KANKRO","BALENO","DOFINO","AGLO","ARKITEKTO","FARMISTO","AGRIKULTURISTO","FISKAPTISTO","B
AKISTO","BUCISTO","KUIRISTO","KELNERO","GASTIGISTO","KOMERCISTO","JURNALISTO","FOTOGRAFISTO","AK
TORO","KANTISTO","DANCISTO","SPORTISTO","TREJNISTO","PILOTO","MARISTO","SOFORO","KONDUKTORO","

MEKANIKISTO","ELEKTRIKISTO","CARPENTISTO","MASONISTO","PENTRISTO","TAJLORO","FRIZISTO","GARDIST
O","TREZORO","BUDGETO","KUSTO","GAJNO","SALAJRO","PENSIONO","ASEKURO","PRUNTO","DEBO","KREDIT
O","RICAJO","MALRICECO","ADMINISTRACIO","MINISTERIO","AMBASADO","KONSULEJO","PASVORTO","PERM
ESILO","LICENCO","PATENTO","AKUZO","PROCESO","ATESTANTO","JUGEJO","PUNAJO","ESCEPTO","REGULO","
SPIRITO","ANIMO","INTELIGENTO","MEMORO","FANTAZIO","ATENTO","KONSCIO","KONSCIENSO","KONVINKO
","OPINIO","KREDO","DUBO","OPTIMISMO","PESIMISMO","MELANKOLIO","SOLECO","NOSTALGIO","KURIOZEC
O","AMBICIO","FIERECO","HONTO","KULPO","ENVIO","JEALUZO","TIMIDECO","MALGAJO","GAJAJO","BRUSTO"
,"VENTRO","DORSO","SULTRO","GENUO","KRURO","TUNO","TUNELO","FLUGO","PROMENO","EKSKURSO","EK
SPEDICIO","AVENTURO","MIGRADO","LIMIGO","PLANO","PROJEKTO","CELO","RIMEDO","REZULTO","FAKO","F
AKULO","MAJSTRO","METIISTO","KOMIZISTO","SEKRETARIO","DIREKTORO","PREZIDANTO","GVIDANTO","ANO
","KOLEGO","LERNEJO","UNIVERSITATO","KLASO","KURSO","LECIONO","EKZAMENO","LETERO","VORTBUTIKO"
,"BIBLIOTEKO","SCIENCO","TEORIO","METODO","EXPERIMENTO","ESPLORO","HISTORIO","GEOGRAFIO","MATE
MATIKO","FIZIKO","KEMIOLOGIO","BIOLOGIO","MEDICINO","TEKNIKO","TEKNOLOGIO","INGENIERO","LABORA
TORIO","DOKUMENTO","DOSIERO","KOPIO","ATESTO","DIPLOMO","KORPO","KAPO","VIZAGO","OKULO","NAZ
O","BUSO","DENTO","ORELO","HARARO","KOLO","BRAKO","MANO","FINGRO","PIEDO","KORO","SANGO","OST
O","HAUTO","MALSANO","FEBRO","DOLORO","KURACISTO","HOSPITALO","MEDIKAMENTO","KURACADO","SA
NO","VIVOENERGIO","DORMEMA","DORMO","REVUJO","EMOCIO","GOJO","TRISTECO","KOLERO","TIMO","ES
PERO","KONFIDO","KURAGO","PACIENCO","BONVOLO","AUTO","BUSO","TRAJNO","STACIDOMO","STACIO","A
VIADILO","FLUGHAVENO","SIPO","HAVENO","BICIKLO","MOTOKIRO","VETURILO","BILETETO","PASPORTO","VA
LIZO","VOJAGO","TURISTO","GVIDLIBRO","MAPO","KOMPASO","MOTORO","RADO","SRAUBO","NAJLO","MAR
TELO","SEGILO","CIFERO","KOMPUTILO","EKRANO","KLAVARO","MONO","MONUJO","BANKO","KREDITKARTO"
,"KONTO","FAKTURO","BUTIKO","MERKATO","KOMERCO","INTERSANGO","PROFITO","IMPOSTO","KONTRAKT
O","OFICO","OFICEJO","FAKO","REKLAMO","KONKURSO","POLITIKO","PARTIO","PARLAMENTANO","MINISTRO
","LEGO","PUNO","KRIMO","MALLIBEREJO","SOCIETO","ASOCIO","KLUBO","KUNSIDO","MUZIKO","KANTO","K
ONCERTO","TEATRO","KINEJO","FILMO","FOTOGRAFIO","PENTRARTO","SKULPTAJO","EKSPOZICIO","SPORTO",
"MATCO","LUDO","KURO","NAGO","PILKO","FERIO","FESTO","INVITO","DONACO","RADIO","TELEVIDILO","JUR
NALO","GAZETO","ARTIKOLO","INFORMO","MESAGO","INTERRETO","RETADRESO","PASATEMPO","FLORARO",
"FRUKTOGARDENO","LEGOMO","KRUCO","KULERETO","SURTUKO","SUBVESTO","NAZTUKO","MANUMBRO","T
USCO","KARTO","KOVERTO","STAMPO","SKRIBOTABLO","BRETO","TIRKESTO","KUSENETO","LITOCHAMBRO","
MANGOCAMBRO","BANOCAMBRO","KELO","ATRIO","BALKONO","TERASO","KORTO","BARILO","PORDEGO","F
LANKACETO","PLACO","PARKO","FONTANO","MONUMENTO","STATUO","PONTO","KANALO","HAVENETO","KA
JERO","LERNOLIBRO","BIBLIOTEKISTO","FAKULTATO","HOMO","VIRO","VIRINO","INFANO","PATRO","PATRINO
","FRATO","FRATINO","AMIKO","MALAMIKO","EDZO","EDZINO","FILO","FILINO","DOKTORO","INSTRUISTO","L
ABORISTO","LERNEJANO","REGO","CIVITANO","AVINO","AVO","ONKLO","ONKLINO","KUZO","KUZINO","NAJBA
RO","GASTIGO","GASTO","MASTRO","SERVISTO","SOLDATO","POLICANO","JUGISTO","ADVOKATO","SCIENCIST
O","ARTISTO","VERKISTO","POETO","MUZIKISTO","AKVO","TERO","SUNO","LUNO","CIELO","AERO","FAJRO","A
RBO","FLORO","HERBO","MONTO","RIVERO","MARO","PLUVO","VUNDO","STONO","BESTO","BIRDO","FISO","
MONDO","ARBARO","LAGO","INSULO","VALO","KAMPO","STELO","NUBO","NEGO","GLACIO","SABLO","BOVIN
O","KATO","HUNDO","CEVALO","PORKO","SAFINO","LEONOJ","URSO","SERPENTO","INSEKTAJ","DOMO","CAM
BRO","PORDO","FENESTRO","TABLO","SEGO","LITO","VESTO","SUO","CAPELO","PAPERO","PLUMO","LIBRO","T
ELEFONO","HORLOGO","GVIDILO","BILDO","LUDILO","ILO","SLOSILO","TEGMENTO","MURO","GRUNDO","TAPI
SO","LUMO","LAMPO","SPEGULO","GARDROBE","KUSENO","KOVRILO","PANTALONO","CEMIZO","VESTAJO","B
OTO","GANTO","MANTELO","KRAVATO","RINGO","CENO","SAKETO","PANO","LAKTO","VIANDO","FRUKTO","V
EGETAJO","SUKERO","SALO","KAFO","TEO","VINO","OVO","POMO","TASO","TELERO","KULERO","FORKO","TRA
NCILO","MANGAJO","MANGO","TRINKAJO","FROMAGO","BUTERO","FISAJO","PORKOVIANDO","BOVOVIANDO
","SUPO","SALATO","TERPOMO","TOMATO","BANANO","ORANGO","CITRONO","KUKO","BISKVITO","BIERO","A
KVUJO","BOTELO","GLASO","SERVVIDO","KUIREJO","TEMPO","TAGO","NOKTO","MATENO","VESPERO","HORO
","MINUTO","JARO","SEMAJNO","MONATO","LOKO","URBO","LANDO","STRATO","VOJO","VICO","METRO","KI
LO","MONERO","VALORO","SEKUNDO","JARCENTO","TAGMEZO","TAGMEZE","VESPERMANGO","VILAGO","PR
OVINCO","KONTINENTO","OCEANO","LIMO","DISTANCO","RAPIDECO","PEZO","LONGO","LARGO","ALTO","GR
ANDO","KVANTO","NOMBRO","PREZO","ENERGIO","ELEKTRO","ELEKTRIZADO","BATERIA","BATERIO","TUBARO
","KABLO","DRATO","STALO","FERO","KUPRO","ORO","ARGENTO","ALOJO","PLUMBO","STANO","KARBO","NA
FTO","GASO","BRULMATERIALO","PLASTO","KAUCUKO","VITRO","CERAMIKO","BETONO","CEMENTO","BRIKET

O","BRIKO","TEGMENTUJO","SINDIGO","PROPAGANDO","CENSURO","GAZETARO","REDAKCIO","REDAKTORO",
"CENZURISTO","ELDONEJO","ELDONISTO","LIBROVENDISTO","PRESISTO","PRESAJO","AFISO","PROSPEKTO","K
ATALOGO","BROSURO","GAZETARTENO","RECENZO","INTERVJUO","DISKURSO","KONFERENCO","SEMINARIO",
"SIMPOZIO","KONGRESO","KUNVENO","DELEGITO","REPREZENTANTO","AMBASADORO","DIPLOMATO","TRAK
TATO","ALIANCO","ORGANISMO","CELO","HISTO","ORGANO","SKELETO","MUSKOLO","NERVO","GLANDO","H
ORMONO","ENZIMO","VIRUSO","BAKTERIO","MIKROBO","PATOGENO","GENETIKO","GENO","HEREDO","SPECI
O","GENRO","FAMILIO","EKOSISTEMO","BIOMO","BIOSFERO","FAUNO","FLORO","FLORA","VEGETAJARO","AR
BARARO","MARBORDO","DUONINSULO","ISTMO","GIMNASTIKO","ATLETIKO","FUTBALO","TENISO","BASKETB
ALO","VOLEIBALO","RUGBYO","KRIMPOLUDILO","GVIDANTO","CIKLISMO","SKIADO","SKETADO","ALPINISMO",
"SURFADO","VELADO","REMADO","ARKPAFADO","SKERMADO","BOKSADO","LUKTADO","STADIONO","GIMNA
ZIO","NAGEJO","KUREJO","KAMPADEJO","FERIEJO","LUDKARTARO","SAKO","SAKTABULO","SAKFIGURO","SUPE
RBAZARO","VENDEJO","KIOSKO","BAKESKOLO","BAKEJO","BUCEJO","SUKERAJO","FARMACIO","FRIZISTEJO","L
AUNDRY","LAVEJO","RIPARISTEJO","GASSTACIO","PARKEJO","POSTEJO","POLICEJO","FAJROBRIGADEJO","URB
ODOMO","KORTUMO","MUZEO","GALERIO","BIBLIOTEKO","ARKIVO","MONUMENTO","KASTELO","PALACO","
MONAHEJO","PREGEJO","KATEDRALO","MOSKEO","KALDRONETO","FORNUJO","KRIBRILO","RASPILO","TIRILO"
,"BOTELMALFERMILO","KONSERVUJO","SKATOLETO","PLATANETO","GLASO","KULEREGO","FORLEGO","LEGAD
O","SKRIBADO","KALKULILO","KAJERO","KALKULTABELO","PRESILO","SKANILO","FOTOKOPIILO","KAMERAO","F
OTILO","MIKROFONO","LAUTPAROLILO","AUSKULTILO","AMPLIFILO","STOPILO","ADAPTILO","KUIRILARO","LU
MTURO","REMPARO","FORTIKAJO","TURTO","NAJTINGALO","GORGULO","STRIGO","FALKO","STORKO","ANSER
O","ANASO","PAVOS","PAVOSO","KOKO","KOKINO","PORKIDO","BOVIDO","SAFIDO","HUNDIDO","KATIDO"};

Ro Pour le ro, un nettoyage des blancs en fin de chaque lemme s'impose sinon le comptage va être biaisé.

ro=StringTrim[{"DUBA ","DUBAL ","DUBAF ","DUBAS ","DUGAB ","DUGAC ","DUGAF ","DUGAL ","DUFAD
","DUFAF ","DUDAF ","DUDAR ","DUDAS ","DUFAS ","DUFAP ","DULAB ","DULAC ","DULAM ","DAMOBA
","DAGOCA ","DAGOFA ","DAGOMA ","DUCAD ","DUCAF ","DUCAK ","DUCAL ","DUJAB ","DUJAF ","DUJAP
","DUJAR ","DUKAD ","DUKAF ","DUKAL ","DUKAR ","DUMAB ","DUMAF ","DUMAK ","DUPAB ","DUPAC
","DUPAL ","DUPAM ","DURAB ","DURAL ","DURAV ","DUTAB ","DUTAC ","DUTAK ","DASAM ","DAJAB
","DAJAD ","DAJAK ","DAJAL ","DAKAB ","DAKAC ","DAKAD ","DAKAF ","DAKAL ","DAKAM ","DAKAP ","DAKAQ
","DAKAR ","DAKAS ","DAKAV ","DALOBA ","DALOCA ","DALODA ","DALOGA ","DALOKA ","DALOMA ","DALOPA
","DALOQA ","DAMOCA ","DAPOBA ","RADAL ","RAFO ","RADAP ","RADAS ","RADAT ","RADAC ","RADAJ
","RADAK ","RADAM ","RADAQ ","RADAR ","RADAPAC ","AR ","RADAW ","RADAY ","RADYAL ","RADNAK
","RADAMAT ","RADAPAV ","BIRA ","BISA ","BISU ","BITA ","BIVA ","BIZA ","BIXAD ","BIRAB ","BIRAC ","BIBA
","BIFAB ","BIDA ","BIGAL ","BIJAB ","BILAB ","BILAK ","BILAF ","BIMAK ","BIMAM ","BIKAB ","BIKAC ","BICAL
","BILNAB ","BIPAD ","BIPAF ","BOCAP ","BOCAR ","BOCAT ","BOBAC ","BOBNOC ","BOVAB ","BOSAB ","BOTAB
","BUBAC ","BUCAB ","BUCAD ","BUCAF ","BUDAB ","BUDAF ","BUFAB ","BUGAB ","BUGAC ","BUJAB ","BUJAD
","BUKAB ","BULAF ","BULAM ","BUMAB ","BUMAK ","BUMAP ","BUMAQ ","BUMAR ","BUMAS ","BUNAD
","BUPAC ","BUPAF ","BUPAL ","BUPAR ","BUPAT ","BUQ ","BUR ","BUS ","BUT ","BUV ","BOKAB ","BOGAB
","BOGOB ","BOJOB ","BOCOD ","BOCNOD ","BOCAL ","BOCAM ","BEGOB ","BEGUC ","BEGUF ","BEGOL
","BEGAP ","BEJAB ","BELUB ","BEMOB ","BEPOB ","BEQOB ","BEROB ","BESUB ","BESNUB ","BEBAF ","BEBAK
","BEBAM ","BEBAR ","BECAB ","BECAF ","BECAL ","BECAP ","BEDA ","BEDAC ","BEDAG ","BEDUK ","BEDUL
","BEDUM ","BEDUP ","BEDUR ","BEDUT ","BEDUV ","BEFAB ","BEFLAG ","BEFLAK ","BEFLAL ","BEFPAD
","BEFAS ","BALAB ","BALAC ","BALAD ","BALAK ","BALAM ","BACAF ","BACAL ","BACAM ","BACAN ","BAFAB
","BAFAD ","BAFAL ","BAFAM ","CA ","CAF ","CAD ","CAK ","CAL ","CAMAB ","CAMIB ","CAMNOB ","CAPIB
","CAQIB ","CARIB ","CASIB ","CATIB ","CEB ","CEDAB ","CEFAB ","CELOB ","CIBUC ","CIBNOC ","CID ","CINAD
","CIFOB ","CIFNOB ","CIK ","CIKNA ","CIL ","CILNA ","CIM ","CIMNA ","CIP ","CIPNA ","CIQ ","CIQNA ","CIRUB
","CIRNUB ","COBUC ","COBNUC ","COD ","CODNA ","COF ","COK ","COL ","COM ","COR ","CORNU ","CUB
","CUBNA ","CUDOB ","CUF ","CUFNO ","CUGOB ","CUGNOB ","CUKOB ","CUL ","CULNU ","CUM ","CUP ","CUQ
","CURUB ","CURNUB ","DUSAB ","DUSAT ","DASAK ","DASAL ","DASAF ","DASAB ","DASAD ","DARAD ","DARAF
","DARAG ","DARAL ","DARAM ","DARAS ","DAMIBA ","DAMICA ","DAMIGA ","DAMILA ","DAMISA ","DAMIDA
","DAMIJA ","DAMIPA ","DAMITA ","DALICA ","DALIDA ","DAPIBA ","RAPIBA ","DAMODA ","DAMIFAB

","DAMOGA ","POJAG ","BOLOF ","BOLOK ","BOLOM ","BOLOS ","BOLAB ","BOLOC ","BEDAF ","PIBWAD ","EBA ","DAGIBEB ","DAGICEB ","DAGILEB ","DAGIMEB ","DABIMEB ","DABIREB ","DABISEB ","DAKITEB "}]

{"DUBA","DUBAL","DUBAF","DUBAS","DUGAB","DUGAC","DUGAF","DUGAL","DUFAD","DUFAF","DUDAF","DUDAR","DUDAS","DUFAS","DUFAP","DULAB","DULAC","DULAM","DAMOBA","DAGOCA","DAGOFA","DAGOMA","DUCAD","DUCAF","DUCAK","DUCAL","DUJAB","DUJAF","DUJAP","DUJAR","DUKAD","DUKAF","DUKAL","DUKAR","DUMAB","DUMAF","DUMAK","DUPAB","DUPAC","DUPAL","DUPAM","DURAB","DURAL","DURAV","DUTAB","DUTAC","DUTAK","DASAM","DAJAB","DAJAD","DAJAK","DAJAL","DAKAB","DAKAC","DAKAD","DAKAF","DAKAL","DAKAM","DAKAP","DAKAQ","DAKAR","DAKAS","DAKAV","DALOBA","DALOCA","DALODA","DALOGA","DALOKA","DALOMA","DALOPA","DALOQA","DAMOCA","DAPOBA","RADAL","RAFO","RADAP","RADAS","RADAT","RADAC","RADAJ","RADAK","RADAM","RADAQ","RADAR","RADAPAC","AR","RADAW","RADAY","RADYAL","RADNAK","RADAMAT","RADAPAV","BIRA","BISA","BISU","BITA","BIVA","BIZA","BIXAD","BIRAB","BIRAC","BIBA","BIFAB","BIDA","BIGAL","BIJAB","BILAB","BILAK","BILAF","BIMAK","BIMAM","BIKAB","BIKAC","BICAL","BILNAB","BIPAD","BIPAF","BOCAP","BOCAR","BOCAT","BOBAC","BOBNOC","BOVAB","BOSAB","BOTAB","BUBAC","BUCAB","BUCAD","BUCAF","BUDAB","BUDAF","BUFAB","BUGAB","BUGAC","BUJAB","BUJAD","BUKAB","BULAF","BULAM","BUMAB","BUMAK","BUMAP","BUMAQ","BUMAR","BUMAS","BUNAD","BUPAC","BUPAF","BUPAL","BUPAR","BUPAT","BUQ","BUR","BUS","BUT","BUV","BOKAB","BOGAB","BOGOB","BOJOB","BOCOD","BOCNOD","BOCAL","BOCAM","BEGOB","BEGUC","BEGUF","BEGOL","BEGAP","BEJAB","BELUB","BEMOB","BEPOB","BEQOB","BEROB","BESUB","BESNUB","BEBAF","BEBAK","BEBAM","BEBAR","BECAB","BECAF","BECAL","BECAP","BEDA","BEDAC","BEDAG","BEDUK","BEDUL","BEDUM","BEDUP","BEDUR","BEDUT","BEDUV","BEFAB","BEFLAG","BEFLAK","BEFLAL","BEFPAD","BEFAS","BALAB","BALAC","BALAD","BALAK","BALAM","BACAF","BACAL","BACAM","BACAN","BAFAB","BAFAD","BAFAL","BAFAM","CA","CAF","CAD","CAK","CAL","CAMAB","CAMIB","CAMNOB","CAPIB","CAQIB","CARIB","CASIB","CATIB","CEB","CEDAB","CEFAB","CELOB","CIBUC","CIBNOC","CID","CINAD","CIFOB","CIFNOB","CIK","CIKNA","CIL","CILNA","CIM","CIMNA","CIP","CIPNA","CIQ","CIQNA","CIRUB","CIRNUB","COBUC","COBNUC","COD","CODNA","COF","COK","COL","COM","COR","CORNU","CUB","CUBNA","CUDOB","CUF","CUFNO","CUGOB","CUGNOB","CUKOB","CUL","CULNU","CUM","CUP","CUQ","CURUB","CURNUB","DUSAB","DUSAT","DASAK","DASAL","DASAF","DASAB","DASAD","DARAD","DARAF","DARAG","DARAL","DARAM","DARAS","DAMIBA","DAMICA","DAMIGA","DAMILA","DAMISA","DAMIDA","DAMIJA","DAMIPA","DAMITA","DALICA","DALIDA","DAPIBA","RAPIBA","DAMODA","DAMIFAB","DAMOGA","POJAG","BOLOF","BOLOK","BOLOM","BOLOS","BOLAB","BOLOC","BEDAF","PIBWAD","EBA","DAGIBEB","DAGICEB","DAGILEB","DAGIMEB","DABIMEB","DABIREB","DABISEB","DAKITEB"}

Lexique Ummite

Le lexique U952 suivant provient de BD VOC qui fournit aussi des traductions ou explications. L'ensemble est proposé par AJH sur Ummo-sciences.org. Il concerne exclusivement les années madrilènes de 1966 à 1969. Il comprend 952 lemmes comme expliqué dans le texte.

U952={"AA","AADAAADA","AADAGIOOU","AADOAUGOO","AADOO","AAGA","AAIIODI","AAIIOYA","AAIODI","AAIODIWOA","AAIOODI","AAKSBOUTZ","AALAADAA","AALOA","AAO","AAR","AARBI","AARBII","AARGA","AARGIAGOO","AARUNNIOGOIA","AARWIO","AASE","AASEE","AASNEII","AASNEUIDAA","AASNOOSAI","AAXAA","AAXOO","AAXOOIBOOA","AAYA","AAYIYAA","ABEE","ADA","ADIUU","ADOGOOI","ADOUAXOO","AEEWWIEE","AELEWEE","AEVO","AEXOOE","AGAXAAU","AGIADAA","AGIIOA","AGIOA","AGIOAA","AGIOOA","AGOIEE","AGOO","AGOODAA","AGOYEE","AIAIEDUNNEII","AIAIEDUNNII","AIDAA","AIGAEGAA","AIGOOXAADA","AIIODI","AIIOODI","AIIOOYA","AIMMOA","AINNAOXOO","AIOBII","AIODI","AIOODI","AIOOIAO","AIOOIEDOO","AIOOYA","AIOOYAA","AIOOYAAIODI","AIOOYAU","AIOOYEDOO","AIOOYEEDOO","AIOYAYEDOO","AMIE","AMIEE","AMMIEYIISAIA","AMMIOXOO","ANAUANAA","ANAUGAA","ANII","ANNOIUXII","AO","AOGAAU","AOUIAOEMII","AOXIBOO","AUAIIAUDOO","AUAIIAUOO","AUGOOA","AUWOA","AWEE","AXOO","AYAA","AYAGGAA","AYII","AYIIO","AYIIYAA","AYIYAA","AYOUUDA","AYUBAA","AYUBAAEWA","AYUBAAEWAA","AYUGAA","AYUU","AYUUBAA","AYUUU","AYUYISAA","BAA","BAAAUAWE","BAAIGO","BAAIGOO","BAAYIODDUU","BAAYIODIXAA","BAAYIODOOUII","BAAYIODOUII","BAAYIODOVII","BAAYIODUIII","BAAYIODUU","BAAYIODUULAA","BAAYIOODUU","BAYIODUII","BEEO","BIAEEI","BIAEI","BIAEMOOXEA","BIAGOO","BIAMOAXII","BIE","BIEE","BIEEBUTSOA","BIEEE","BIEEGOO","BIEEUNNIEO","BIEEWIA","BIEEWIGUU","BIEUIGUU","BIEYAE","BIEYAEYUEO","BIIEUIGUU","BIIO","BIUA",

"BIUUA","BOIYUIYOE","BOIYUIYU","BOO","BOOGOO","BUAAWA","BUAE","BUAWAA","BUAWAAMIESEE","BUA
WAIGAAI","BUAWE","BUAWEE","BUAWWAA","BUEA","BUMAEOEMII","BUUA","BUUAWA","BUUAWAAA","BU
UAWAAMIESEE","BUUAWAM","BUUAWAMIESEE","BUUAWE","BUUAWE","BUUAWEA","BUUAWEE","BUUAW
OEMII","BUUAWWA","BUUTZ","BUUTZGOIA","BUUXIIAO","CSI","DAA","DAII","DAO","DAXEE","DEE","DIEWEE",
"DII","DIIO","DIIU","DIIUYA","DIIUYAA","DIUYAA","DNMAAEI","DO","DOA","DOAA","DOAWAA","DOEE","DOO"
,"DOOXOO","DOOYUIIABII","DOROO","DOROOUIAAIE","DOUIAMOO","DU","DUII","DUIO","DU-OI-
OIYOO","DU-OI-
OIYOO","DUU","DUUO","EAAIODI","EABAYO","EAYODI","EBAYAA","EBEE","EDAAU","EDDIO","EDDOIBOOI","ED
IEEDI","EDUU","EE","EEAEE","EESEE","EESEEOEMI","EESEOEMI","EESEOEMII","EESEOMI","EEUE","EEWE","EEW
EANIXOO","EEWEE","EEWEEANIXOO","EEWEEGOO","EI","EIBIENEE","EIDIIU","EIMII","EIWOO","EIXAE","EIXUU
A","ELOO","ELOOWA","EMOO","ENMEE","ENMMOO","ENMO","ENMOO","ENNAEOI","ENNAEOII","ENNAOEI",
"ENNAOEI","ENNAOEII","ENNOI","ENNOII","EOEEI","ESEE","EUNNA","EUUNNA","EWOIAAUIU","EXAABI","EXA
ABII","EXXAABII","EYAOLOOWA","EYAOLOWAO","EYEA","GAA","GAADII","GAARAADUI","GAEOAO","GEE","GIA
A","GIADAA","GIAXAA","GIIA","GIIDII","GIIUULOO","GIIXAA","GIUDUUDAA","GOA","GOAA","GOABAAAE","GO
AWAA","GOO","GOOA","GOOAIE","GOODAA","GOONIIOADOO","GOONNIAOADOO","GOSEEE","GUU","IA","IA
AIYAA","IAAS","IADDUU","IADUU","IAGAIAAOO","IAGIAAIAOO","IAI","IAIKEAI","IAIKEAIUUXAA","IANNO","IAN
NOO","IAOOI","IAS","IASAABII","IASXOODINAA","IAVAMII","IAWA","IAWIAIA","IAXAABI","IAXAABII","IBOAAAY
A","IBOAAIA","IBOAAYA","IBOAAYANOA","IBOAAYANUUIO","IBOAYA","IBOAYAA","IBONEE","IBOO","IBOOAYA
","IBOONEE","IBOTZOOIXINAA","IBOXOODINAA","IBOZOO","IDDOANAA","IDIA","IDIIA","IDUGOO","IDUIROO","
IDUUWII","IDUUWIIO","IDUWII","IE","IEAAYA","IEBOO","IEE","IEES","IEGOOSAAA","IEGOSAA","IEN","IENXOOD
INAA","IES","IEVOOXOODINAA","IEVOXODINAA","IEYIOBAA","IGAAYUU","IGAYUU","IGIAAYUYISAA","IGIO","IGI
OI","IGOOA","IGOOMII","IGOONOOI","IGUU","IGUUXOO","IIA","IIAS","IIE","IIO","IIOSSAUUIGAA","IISUIW","IIT
OA","IITOOA","IIWEEWEE","IIWOAE","IMAAUIII","IMMAA","IMMI","INAIE","INAYUYISAA","INNAYUYISAA","INN
UO","INNUOVIAAXOO","INOWII","IOAOOI","IOAWO","IOAWOO","IOAXOO","IOGAARAA","IOGIOEMMII","IOIX
OINOIYAA","IOSAAXII","IOUMMA","ISIAGEE","ISOO","ITOAA","ITOOAA","IUAGAROO","IUAMMIO","IUDIIXAA","
IUMA","IUMMA","IUMMAA","IUMMASNEII","IUOOROO","IVOAAIA","IVOOROO","IWA","IWO","IWOIA","IWOU
AXOO","IXAABII","IXIMOO","IXINAA","IXIOGAA","IXOIAROO","IXOOURAA","IXOUURAA","IXWEE","IYOAEE","KA
AWAEA","KEAI","KEAIA","KEEAI","KEOYEEOO","KOAE","KOOAE","LAAIYAA","MIIBAYEAAO","NA","NAA","NAAN
AA","NAATOWSEE","NAAU","NAAXUNII","NAE","NAI","NAII","NAUEE","NEAAA","NIAAIODOUI","NII","NIIAXOO
","NIIO","NIIOADOO","NIIUAXOO","NOA","NOAUIW","NOAXUNII","NOI","NOIA","NOOLAWE","NOOXOEOOYAA
","NOURAXAA","NUAEL","NUGII","NUIO","NUUDAIAA","NUUGI","NUUGII","NUUYAA","OA","OAA","OACAWA",
"OAE","OAG","OAGOOU","OAI","OAK","OAKEDEEI","OANA","OANAUAA","OANEEA","OANEEAOIYOOYO","OAN
MA","OANMAA","OANNEAOIYOYOO","OAROO","OASION","OAUOLEA","OAUOLEEA","OAUOOELEA","OAWOEN
NIU","OAWOENNIUU","OAWOLEA","OAWOLEA","OAWOLEAIDA","OAWOLEAIDAA","OAWOLEIIDAA","OAWOO
","OAWOOALEAIDAA","OAWOOALEIDAA","OAWOOENII","OAWOOLEA","OAWOOLEAIDAA","OAWOOLEEA","O
AWOOLEIBOZOO","OAWOOLEIDAA","OAWOOYEY","OAWUO","OAWW","OAXAUXAA","OAXOO","OAYU","OAY
UU","OBEE","OBUANAA","OBXANWAII","ODAWAA","ODAXIXAA","ODEU","ODIAA","ODU","OEBUMAEI","OEBU
MAEOEEMII","OEBUMAEOEEMIII","OEBUMAEOEMII","OEBUMAOEMII","OEE","OEEMII","OEEMM","OEI","OEII"
,"OEM","OEMBUAW","OEMBUAWW","OEMBUUAM","OEMBUUAN","OEMBUUAW","OEMI","OEMII","OEMIIAB
II","OEMIIGIIA","OEMIII","OEMIIUEWA","OEMM","OEMMI","OEMMII","OEMMIIUEWA","OEMMIOYAGAA","OE
NBUUAU","OERUU","OES","OESEE","OEUDEEE","OEW","OEWEA","OGIA","OGIAA","OGIIA","OGOKOOA","OI","
OIAGAA","OIBIIA","OII","OINNA","OIWI","OIXII","OIXIIXI","OIXXIIXI","OIYOO","OIYOYOIDAA","OLEASS","OMAIU
","OMGEEIIE","OMGEEYIE","OMGEYIE","OMGGEEYIE","OMMOI","OMWAA","OMWAA","OMWEA","ONAAWO",
"ONAUDO","ONAUDOOXA","ONAUUU","ONAWO","ONMAEI","ONNOXOO","OOAA","OOAWOO","OOBO","OO
BOO","OOE","OOGIXUA","OOGIXUAA","OOGOOIAA","OOL","OOLEA","OOLEEA","OOLGA","OOLGAA","OOMO",
"OORGAOWI","OOU","OOXAII","OOXEE","OOXENNUU","OOYAA","OOYAAUYIEE","OOYAEBEEM","OOYAGAA","
OOYAGAAWOA","OOYAUMME","OOYAUNMEEEI","OOYAWIIA","OOYIA","OOYIAA","OREEAU","OU","OUDEE","
OUDEXIENOO","OUEWA","OUIA","OUIAOEM","OUIAOEMII","OUIW","OULIOOA","OUMBOO","OUMBOOBUUA
","OUMBOOMIIA","OUMDAA","OUMMO","OUMMOAELAEWE","OUMWI","OUMYASAAII","OUSEIOSAA","OUU
","OWEEEU","OWEEU","OWOALEAIDA","OWOOLEAIDAA","OXA","OXOEEOIADUU","OXOEOIADUU","OXOIAEE",
"OXOOIAEE","OXOOOIAEE","OXUO","OYAA","OYAAAGAA","OYAAGA","OYAAGAA","OYAGAA","OYISAA","OYUU
","SAABI","SAAOA","SAAOOA","SANMOO","SAWA","SI","SIAA","SIUU","TAAU","TAS","TAU","TAXEE","TERROE
WAEN","TOOKAAIA","UA","UAA","UAAAM","UAAM","UAAYUBAA","UAEXOOE","UALEEXII","UALI","UAM","UA

MII","UAMIIGODAA","UAMIIGOODAA","UAMIIGOOINUU","UAMIIOWOBO","UAMIIXAABII","UAMIIXANMOO","UAMM","UAMMIXANMOO","UAXAOO","UAXOO","UAXOOAXOO","UAXOOGOINNUU","UBAA","UBOO","UDIOO","UDIXAA","UE","UEIN","UEUA","UEUAA","UEWA","UEWAA","UGEE","UGEEYIE","UGII","UGUUAXIIA","UI","UIAA","UII","UIIOGAA","UIIORAA","UIMMA","UIW","UIWIIOO","UIWIUTAA","ULAAYANA","ULAAYANANAE","ULNII","UMMO","UMMOAA","UMMOAAELEUE","UMMOAELEVEANI","UMMOAELEWE","UMMOAELEWEANII","UMMOAELEWEE","UMMOAELEWEEANNI","UMMOALEWE","UMMOALEWEANNI","UMMOALEWEEANNI","UMMOAOO","UMMOELEWE","UMMOELEWEE","UMMOEMII","UMMOEMMII","UMMOGAEOAO","UMMOGAIA","UMMOGAIAO","UMMOGAIAOO","UMMOGAOEAAO","UMMOOEMII","UMMOTAEEDA","UMMOWOA","UMOOEMII","UMYAE","UNAUO","UNAUOWI","UNAWO","UNIEYAA","UNIIEAOEMII","UNIIEAYAAEBU","UNNIAXII","UNNIEAOEMII","UNNIEYA","UNNIEYAA","UNNIEYO","UNNIEYUU","UNNIOOGOIA","UO","UOAD","UOOAMAII","UOUAMII","UOUORAA","UOUOURAA","UOXA","UOXODINAA","UOXOODINAA","UOXOODINNAA","URAA","URAAIWOO","URAIWO","URIEWOO","USAAGIXOO","USADAADAU","USAGIISOO","USDOUOO","USGIGIIAM","USII","USUIW","UUAA","UUDEEXAA","UUDIII","UUDUAGOO","UUDUINOO","UUEIN","UUGEE","UUGEEYIE","UUI","UUIDDAO","UUIORAA","UUL","UULAXAA","UULAXBOIYU","UULAYA","UULAYANA","UULEEWUAA","UULEWUA","UULIBOO","UULIBOOA","UULIXINAA","UULNII","UULODAASAABII","UULODASAABII","UULODAXAABI","UULODOO","UULOODOO","UULUAXOO","UULUEUUA","UULUEWA","UULUEWAA","UULUEWUA","UULUEWUAA","UULUEWUUA","UULWA","UULWAAGAA","UULXOODII","UUUDEESA","UUWAAM","UUWUUA","UUXAEEMOI","UUXAGIXOO","UUXAO","UUYABOO","UUYI","UUYIE","UWAAM","UWAAN","UWAMM","UWAN","UWOSS","UWUUA","UWWAN","UXGIGIAM","UXGIGII","UXGIGIIAM","UXGIGIIAMM","UXGIIGIIAM","UXII","UYAAAYAA","UYI","UYIE","UYIEXAA","UYII","UYIIE","UYOAA","UYOALAADAA","UYOOXIGEE","UYUUNOODII","VAAUAWE","VAUAWEE","VIAAXOO","VIW","VOAT","WAALI","WAAM","WAAMDISAIAYA","WAAMIAAYO","WAAMOO","WAAMTOA","WAAMUA","WAAMWAAM","WAAN","WAANWAAM","WADDOSOOIA","WADOOXOAA","WADOOXOIA","WAEELE","WAEELEWIE","WAELEUIE","WALIOA","WE","WEAM","WEE","WI","WIE","WIISIIO","WIIWIIAA","WIOA","WOA","WOABAEEYUEE","WOAIIOO","WOALA","WOALAOLEASS","WOALAOLOO","WOASSEE","WOAT","WOIOA","WOIWOAXAABI","WOIWOI","WOIWOIXAABI","WOO","WOODO","WOODOO","WUA","WUAM","WUANAI","WUUA","WUUA","WUUNUA","XAABI","XAABII","XAABIUANAA","XAABIUANNAA","XAANIBOOA","XAANMOO","XAAXADO","XAAXADOO","XAIUU","XAN","XANIBOA","XANIBOOA","XANMMO","XANMMOO","XANMO","XANMOO","XANMOULAYA","XANMOUULAYA","XANWAABUASII","XANWAABUUASII","XAUIW","XAXOOU","XEE","XEEUMMO","XI","XII","XIIXIA","XIIXIOUIA","XIIXIOUIAA","XINIBOOA","XOA","XODINAA","XOIMMA","XOODI","XOODINAA","XOODINAADOO","XOODIUMMO","XOODIUMMOO","XOOGU","XOOGUU","XOOIMAA","XUANOO","XUU","YAA","YAA","YAAEDINNOO","YAAESIOO","YAAIODEE","YAAIOSU","YAAIOSUU","YAAXAIIU","YAAYODEE","YAEDINNOO","YAEDINOO","YAEYUEYEDOO","YEDDO","YIE","YIEE","YIEKEAI","YIIE","YIIEAGAA","YIISAIA","YIISA-
OO","YOAAGAA","YOAXAA","YODEE","YOEIM","YONNIANA","YONNIANNA","YONNIANNAA","YOOXAO","YOYGOAAXOO","YUEO","YUMMA","YUUWO","YUUXAA","YUUXIIO","YUUXIO","YUXAA","YUXIDOO","YUXIIO"};

Comme on le voit, les tailles des différents lexiques comprennent quelques centaines de lemmes et ces tailles sont très loin d'être équivalentes à celles des lexiques complets des différentes langues choisies.

L'objectif était de rester dans le même ordre de grandeur que pour le lexique Ummite des années madrilènes comprenant 952 lemmes. Cela semble suffisant pour les estimations globales du texte.

Map[Length,{U952,esp,FR,ranto,ro,turc,HAW}]

{952,609,600,800,321,751,345}