

## TABLEAUX

Les tableaux présentent différentes informations sur le mot "Orient" dans les langues choisies. Chaque tableau contient plusieurs colonnes :

- La colonne Code HTTP indique le résultat d'une requête web, si la page a été trouvée avec succès, cela affichera (200) ou si au contraire, il y a eu une erreur, cela indiquera (4xx ou 5xx).
- URL est liste les adresses web correspondantes.
- Encodage montre quel format est utilisé pour afficher les pages. Si aucune information n'est indiquée, le navigateur utilise un encodage par défaut.
- Aspirations, sauvegarde une copie complète de chaque page dans un fichier texte, tandis que la colonne Dumps enregistre uniquement le contenu textuel des pages.
- Occurrences affiche combien de fois certains mots apparaissent, selon des expressions régulières.
- La colonne Contexte extrait chaque occurrence de ces mots avec une ligne de texte avant et après, puis sauvegarde ces extraits dans des fichiers texte.
- La colonne Concordances montre chaque occurrence accompagnée des mots situés immédiatement à gauche et à droite.

## ANALYSE LINGUISTIQUE

Le corpus analysé repose sur un ensemble de 30 URL sélectionnées pour explorer le mot « orient » dans différents contextes en espagnol. Ces sites couvrent plusieurs thèmes évoquant l'art, l'éducation, la culture, mais aussi la géographie. Ceux-ci permettent de dégager des catégories d'usage principales, reflétant la polysémie du mot. Il ne s'agit là que d'un exemple représentatif d'un corpus plus vaste, mis en évidence dans le nuage de mots, où l'on peut observer en profondeur tous les termes gravitant autour du mot étudié.

Lors de la recherche des sites, je n'ai pas souhaité me concentrer sur un thème spécifique, mais plutôt sur les premiers résultats affichés par les moteurs de recherche. L'objectif était de voir quels mots étaient mis en avant au moment de chercher le terme « oriental » en espagnol. Par conséquent, l'un de nos intérêts communs est de révéler s'il

existe des différences marquantes en termes de stigmatisation de l'Orient, ainsi que la manière dont il est évoqué dans les hémisphères occidental, oriental et sud.

Tout d'abord, analysons le résultat du script cooccurrent. Pour cela, j'ai utilisé la commande :

```
python3 /Users/catalinaalvarez/cours/plurital/PPE1_groupe_2025/programmes/
cooccurrences.py \
/Users/catalinaalvarez/cours/plurital/PPE1_groupe_2025/PALS/espagnol_pals/dumps-
text-espagnol.txt \ --target '.*orient.*' --match-mode regex \-s i \
> /Users/catalinaalvarez/cours/plurital/PPE1_groupe_2025/PALS/dumps-text.tsv
```

<b>token</b>	<b>corpus size</b>	<b>all contexts size</b>	<b>frequency</b>	<b>co-frequency</b>	<b>specificity</b>
<b>asia</b>	5008	3764	98	97	10.77
<b>medio</b>	5008	3764	55	54	5.60
<b>estudios</b>	5008	3764	28	28	3.48
<b>arte</b>	5008	3764	26	26	3.23
<b>año</b>	5008	3764	26	26	3.23
<b>norte</b>	5008	3764	25	25	3.11
<b>occidente</b>	5008	3764	24	24	2.98
<b>ha</b>	5008	3764	22	22	2.73
<b>discurso</b>	5008	3764	19	19	2.36
<b>más</b>	5008	3764	26	25	2.30
<b>término</b>	5008	3764	18	18	2.24
<b>parte</b>	5008	3764	17	17	2.11
<b>pirineos</b>	5008	3764	17	17	2.11
<b>no</b>	5008	3764	36	33	2.06
<b>fue</b>	5008	3764	16	16	1.99
<b>mando</b>	5008	3764	22	21	1.87
<b>siglo</b>	5008	3764	14	14	1.74
<b>pintura</b>	5008	3764	14	14	1.74
<b>les</b>	5008	3764	14	14	1.74
<b>áfrica</b>	5008	3764	14	14	1.74
<b>su</b>	5008	3764	36	32	1.62
<b>temas</b>	5008	3764	13	13	1.61
<b>es</b>	5008	3764	69	58	1.59
<b>próximo</b>	5008	3764	12	12	1.49
<b>lo</b>	5008	3764	33	29	1.41
<b>qué</b>	5008	3764	10	10	1.24
<b>the</b>	5008	3764	10	10	1.24
<b>españa</b>	5008	3764	10	10	1.24
<b>todo</b>	5008	3764	15	14	1.17
<b>cultura</b>	5008	3764	9	9	1.12

En examinant les concurrents du mot « orient » dans notre corpus, nous remarquons la présence du terme « medio », qui fonctionne ici comme modificateur adjectival du nom « orient ». De même, on trouve « estudios » et « arte », qui jouent le rôle de compléments du nom étudié. Sur le plan syntaxique, nous pouvons donc déduire que « orient » n'est pas seulement défini comme un lieu situé dans l'espace géographique , en l'occurrence « medio oriente », mais aussi comme un cadre englobant des branches académiques et artistiques.

Nous avons d'abord analysé la fréquence globale et la distribution des variantes. Pour ce faire, nous avons ciblé tout texte contenant la séquence « orient », qu'elle apparaisse au début, au milieu ou à la fin de la chaîne : .\*orient.\*. Étant donné que le mot « orient » est vaste et peut prendre plusieurs formes, nous avons cherché à capturer toutes ses occurrences, quelle que soit leur variante. La fréquence totale dans le corpus s'élève à 917 occurrences, réparties comme suit :

Oriente (205)

Orientalismo (193)

Oriental (173)

Orientales (110)

Orientalista (102)

Orientalistas (50)

Orientaliedad (28)

Orientalism (13)

Orientación (11)

Orientalismo (4)

Autres variantes (entre 3 et 1 occurrence).

Nous allons maintenant explorer les cooccurrences les plus spécifiques. La colonne « spécificité » mesure la force de l'association entre l'objet d'étude et ses cooccurrences. Le mot « asia » (spécificité = 10,77) est fortement associé à « oriente » dans un contexte géographique, un sujet d'étude culturellement central. Pour « medio », la spécificité atteint 5,60 : cette association renvoie également à une zone géographique, mais avec une nuance. « Medio » désigne une entité précise, un marqueur politique et international, tandis qu « asia » évoque plutôt un cadre culturel et historique. Enfin, si l'on regroupe les fréquences statistiquement proches, on observe que des mots comme « estudios » (3,48) et « arte » (3,23) reflètent cette dimension culturelle évoquée précédemment. De même,

token	corpus size	all contexts size	frequency	co-frequency	specificity
asia	40694	3912	147	97	61.51
medio	40694	3912	75	54	37.66
pirineos	40694	3912	20	17	14.38
norte	40694	3912	52	25	11.96
áfrica	40694	3912	17	14	11.55
occidente	40694	3912	58	24	9.81
discurso	40694	3912	38	19	9.62
rmino	40694	3912	24	15	9.54
estudios	40694	3912	85	28	8.65
año	40694	3912	39	18	8.45
es	40694	3912	276	58	8.22
les	40694	3912	27	14	7.52
tā	40694	3912	54	20	7.34
ha	40694	3912	66	22	7.06
ximo	40694	3912	10	8	6.57
contemporánea	40694	3912	13	9	6.48
refiere	40694	3912	11	8	6.05
geogrā	40694	3912	11	8	6.05
describir	40694	3912	11	8	6.05
temas	40694	3912	32	13	5.52
prā	40694	3912	20	10	5.35
oldid	40694	3912	5	5	5.09
afganistān	40694	3912	5	5	5.09
pakistān	40694	3912	5	5	5.09
acadā	40694	3912	30	12	5.06
pintura	40694	3912	41	14	4.88
edward	40694	3912	48	15	4.67
pintores	40694	3912	15	8	4.64
hacer	40694	3912	12	7	4.44
perceval	40694	3912	12	7	4.44
tanto	40694	3912	34	12	4.44
mico	40694	3912	25	10	4.32
parte	40694	3912	64	17	4.22
francés	40694	3912	17	8	4.15
said	40694	3912	133	27	4.05

« discurso » (2,11) et « occidente » (2,36) relèvent de la sphère politique à laquelle « medio » fait référence.

L'analyse des contextes permet de valider ou d'affiner nos observations précédemment effectuées sur les dumps. En examinant si les cooccurrences identifiées dans les dumps textuels se retrouvent dans ces contextes. Cela permettra également d'éclairer des usages plus subtils ou des nuances qui échappent à l'analyse basée sur les dumps uniquement. Voici la commande utilisée pour cela : python3 /Users/catalinaalvarez/cours/plurital/PPE1\_groupe\_2025/programmes/cooccurrences.py \ /Users/catalinaalvarez/cours/plurital/PPE1\_groupe\_2025/PALS/espagnol\_pals/contextes-espagnol.txt --target '.\*orient.\*' --match-mode regex \-N 30 \-s i \> /Users/catalinaalvarez/cours/plurital/PPE1\_groupe\_2025/PALS/contextes.tsv

Les cooccurrences fréquentes restent globalement les mêmes que dans les dumps. Les plus spécifiques révèlent les associations immédiates du mot cible : « asia », toujours fortement lié à la notion de légende, tout comme « medio ». En revanche, « pirineos » et « norte » prennent cette fois une place importante, ce qui souligne la forte association de « orient » avec une zone géographique, suscitant davantage d'intérêt pour son ancrage spatial que pour d'autres thèmes. Les mots « áfrica », « occidente » et « discurso » renforcent cette idée de « orient » comme un terme éveillant un vif intérêt géopolitique, souvent central dans le discours politique occidental.

Pour finir, si nous faisons une comparaison des nuages de mots crée d'après la totalité du corpus des dumps textuels et des contextes respectivement, qui reprennent l'ensemble du contenu des 50 pages web. L'analyse visuelle des nuages de mot vient compléter celle réalisée avec les scripts. Alors que les scripts permettent de mesurer la fréquence des mots et leur répartition dans le textes, la visualisation aide à comprendre les grands thèmes et leur importance relative. Dans les dumps, « Asia », « said », « historia », « cultura », « mundo » ressortent visuellement comme dominants, alors que dans les contextes, ce sont « medio », « asia », « estudios » qui prennent le dessus.

Les scripts permettent également d'identifier des statistiques détaillées sur la diversité et la densité lexicales, mais ils n'en révèlent pas directement la nature. Un simple coup d'œil au nuage de mots permet de constater que les thèmes sont plus variés lorsqu'ils font référence à la culture et à l'art, tandis que les contextes se concentrent davantage sur « oriente » en tant que zone géographique. De plus, la taille des mots dans le nuage

indique directement quels sujets dominent dans chaque source, alors que les scripts nécessitent une interprétation des chiffres.

En somme, chaque suffixe nominal associé à « oriente » évoque un point de vue distinct :

« -ismo » désigne un courant esthétique ou intellectuel ;

« -ista » forme le nom de personnes, de spécialistes ou de professions ;

« -ción » renvoie à une action ou un processus.

On peut donc en déduire qu'en espagnol, sur le web, « oriente » est perçu comme une zone d'étude géopolitique et culturelle majeure, suscitant un intérêt marqué non seulement dans les pays occidentaux, mais aussi en Amérique du Sud.