Aliança per la presència digital del català

ANALYSIS OF THE VISIBILITY OF CONTENT IN THE CATALAN LANGUAGE IN INTERNET SEARCH RESULTS

ENGLISH

MAY 2024

ACCIÓ CULTURAL DEL PAÍS VALENCIÀ · AMICS DE LA WIKIPÈDIA · fundació .cat · INSTITUT D'ESTUDIS CATALANS · institut ramon llull · OBRA CULTURAL BALEAR · ÒMNIUM LLENGUA CULTURA PAÍS · PLATAFORMA PER LA LLENGUA · SOFTCATALÀ · WIKIMEDIA CAT

SUMMARY

INTRODUCTION

Since the third quarter of 2022, many Catalan-speaking Internet users have noticed that **the Catalan content had become considerably less visible in web search results** in comparison with the past, even when the user browsing environment was configured to prioritize Catalan content. This meant that, under equal conditions, in searches Catalan was being pushed well below its rightful position in terms of number of speakers.

In view of this anomaly, on March 21, 2023[1], ten entities of civil society formed the Aliança per la Presència Digital del Català (Alliance for the Digital Presence of Catalan), with the urgent, initial milestone being proposed as **to record the scope of the problem** with objective, representative data **in order to address web browser operators** and demand a solution.

Commissioned by the Alliance, the .cat Foundation asked a large number of organizations from the business, academic, public administration, and media sectors operating websites with content in Catalan and at least one other language to provide the log of **traffic data from the web browsers** in each of the languages they offered. Overall, these **contributing organizations operate over 600 websites**, including domains and sub-domains.

On receiving the data, the .cat Foundation analyzed the evolution of visits to the Catalan versions compared with visits to versions in other languages and produced a report to confirm that **traffic in Catalan had fallen on two out of every three websites studied** and, in 80% of the cases, this fallen traffic **had switched to the Spanish version**. The report provided different technical hypotheses about the reasons for the incident, most of which were ruled out, and the conclusion was reached that **the source of the problem lay in a change in** browser classification **algorithms**.

This report was published on June 6, 2023 and can be viewed on the Alliance's website[2]. Alongside this, **the report was sent to the Engineering departments of the two most popular browsers**, Google and Bing, which welcomed it with interest and agreed to look into ways of reversing the problem.

After several months of intense technical dialog with the team at .cat Foundation, **in late August 2023 Google announced a core update** of its browser, the most widely used in the world, to improve how the interests of each user were interpreted. According to the company, this change favored many non-majority languages, although the collaboration of

[1] https://aliançadigital.cat/wp-content/uploads/2023/03/NdP_Alianca-per-la-presencia-digital-del-catala_.pdf

[2] https://xn—alianadigital-mgb.cat/wp-content/uploads/2023/06/informe_posicionament_v.1.11_CAT.pdf

Catalan-speaking users and of the Aliança per la Presència Digital del Català was explicitly[3] welcomed.

Our mechanism to systematically monitor the search results for around twenty random key words instantly **showed an improvement in the positioning of Catalan**[4], which started to exceed Spanish in some terms. Since then, we have kept this mechanism active to check that the improvement has remained steady.

In addition to this monitoring, **we have updated the initial study**, re-analyzing the traffic on hundreds of multilingual websites from browsers. The results of this update and proposed new actions are included in this document.

[3] https://twitter.com/searchliaison/status/1700057703421268374

[4] https://xn--alianadigital-mgb.cat/wp-content/uploads/2023/09/catala-visibilitat.pdf

EXECUTIVE SUMMARY

On commission by the Aliança per la Presència Digital del Català, the .cat Foundation recorded in mid-2023 that content in Catalan was considerably less visible in Internet search results by analyzing the traffic on over 600 contributing multilingual websites. At the time, it was estimated that two out of every three websites during 2022 had less traffic to their Catalan version, and it was seen that the phenomenon even affected Internet users who had their environment configured to prioritize content in Catalan.

The results of the analysis were sent to the Engineering department at Google, which responded by applying a core update in late August, which seemed to ensure the corresponding visibility of the Catalan content was restored.

In early March 2024, **six months after the core update by Google, the .cat Foundation repeated and extended this analysis to quantify the apparent improvement**, re-analyzing the traffic on a similar set of websites that add up to around 400 million browser sessions within the period of almost three years considered, from early 2021.

The study confirms that, based on a level of 41% during the first 11 months of 2021, the relative weight of Catalan content in hits from the Google browser fell below 34% at the end of 2022, and remained around this level for almost eight months in 2023.

After being questioned by the Alliance with the data from the original report, **in late August 2023 Google applied** a core update to ensure the intentions of users were better interpreted, **and the relative weight of Catalan skyrocketed** by 18% to reach 52%. **After this, traffic in Catalan has stabilized** at around 46%, which is a 12% improvement on the worst times of the visibility crisis.

The Alliance plans to continue closely overseeing the visibility of web content in Catalan through different ongoing monitoring activities combined with promotional campaigns for the linguistic normalization of the Catalan language in devices.

METHODOLOGY OF THE STUDY

In this new edition of the study of the positioning of Catalan content in Internet search results, the same methodology as in the original study was applied[5], comparing visits from browsers to the Catalan versions with those to versions in the other languages that each website offers. This report focuses on **traffic from the Google browser**.

In some cases, the team at .cat Foundation was able to directly access the website traffic analysis tools. In others, their owners extracted the data according to our specifications and provided us with them. These specifications can be seen in the annexes to the first edition of the report.

This time, a number of websites similar to that of the original study, **around 600 between domains and sub-domains**, was used. What is more, some less relevant websites were replaced with others with a much greater volume of traffic, increasing the representativity of the results.

Moreover, in most cases **we have had an even broader historical series than in the original report**. In this case, we were able to analyze the evolution of traffic since early 2021—to provide the reference of levels before the incident—until late February 2024—six whole months after the core update by Google.

[5] https://xn—alianadigital-mgb.cat/wp-content/uploads/2023/06/informe_posicionament_v.1.11_CAT.pdf

RESULTS OF THE STUDY

This results section is divided into two parts: first, the decline in traffic in Catalan from Google to the contributing websites experienced after the first positioning study is reviewed. The overall evolution of visits in Catalan to the set of contributing websites is then reviewed, based on a long-time frame which includes the past three years, from January 2021 to February 2024.

DECLINE OF CATALAN IN GOOGLE RESULTS

Over the months after the period reviewed in the first positioning report, there was still a decline in traffic from the Google browser to the Catalan versions of the contributing websites. This period began after the second quarter of 2023 and reached its peak in August of that same year.

Google then acknowledged the problem and applied changes to its system, which reverted the trend and resulted in a notable—or in some cases even sudden—improvement in traffic to the Catalan versions.

To show this situation, **the evolution of visits experienced by three of the collaborators in the study is indicated below**. These include a diagram of the status of traffic between week 27 (from July 3 to 9) and 43 (from October 23 to 29) of 2023.

COLLABORATOR 1

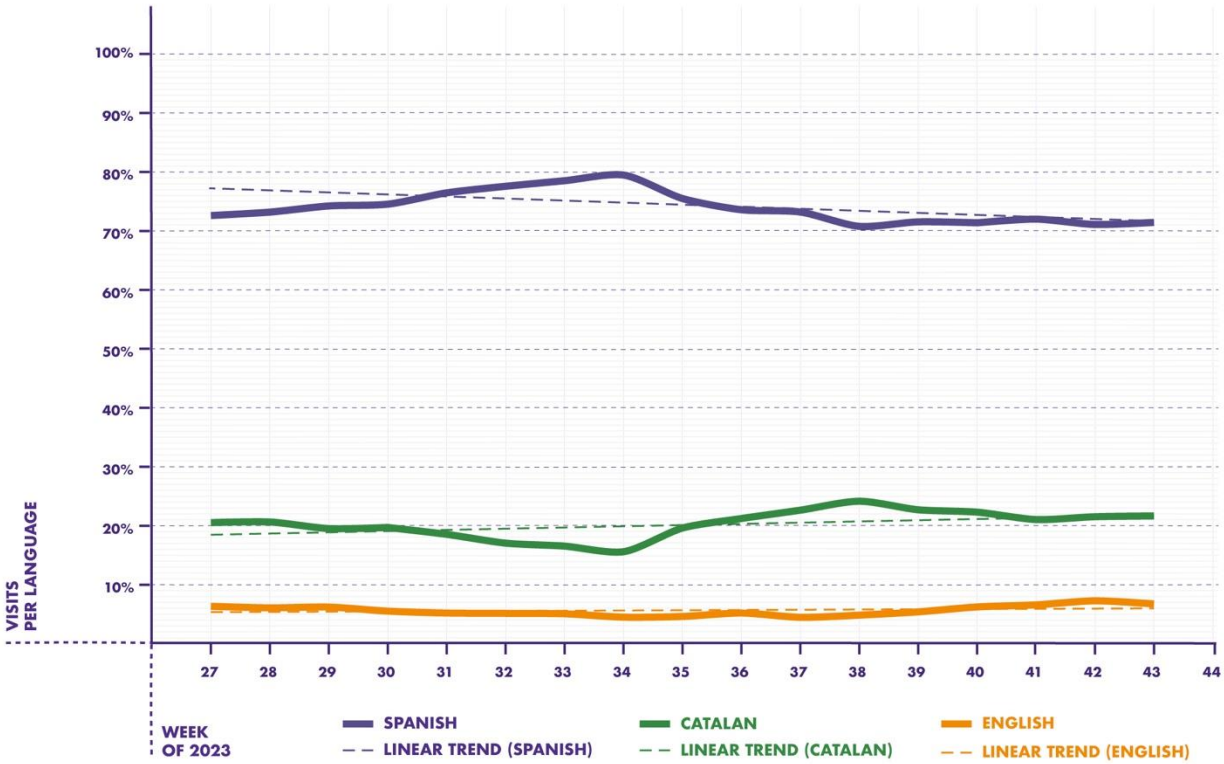
DESCRIPTION

Higher education institution with primarily national students, which includes three languages symmetrically in its web environments.

Correlation CA – ES -0.95
Correlation CA – EN 0.46

Does it use "hreflang"?
Yes, it formats correctly.

Full series
(X-axis: week; Y-axis: percentage of traffic from Google for each language)



The value of traffic to the Catalan version for the first collaborator began at around 21% (average for the first 4 weeks), although this suddenly declined to hit a minimum of 15.68% during week 34. At this same time, the Spanish version reached its peak at around 80% (79.54%), whereas the English version remained at minimum values below 5% (4.79%).

As soon as Google acknowledged the problem and started to apply solutions, the Catalan version began to climb to reach the peak of the series, exceeding 24% (24.07%) in week 38. As in the cases examined in the first positioning report, this is the same time as the lesser presence of the Spanish version, which drops by 8% to stand at 71% (70.91%). The change lies with the English version which, after a very mild recovery, dropped again in week 37 to stand at barely 5% (5.01%).

Comparing the average of the first 4 weeks with that of the last 4, the Catalan version was able to end the series with a rise of 1.3%, primarily due to the 1.2% loss of the Spanish version. However, the English version ended with the same initial figure of 6.8%.

COLLABORATOR 4

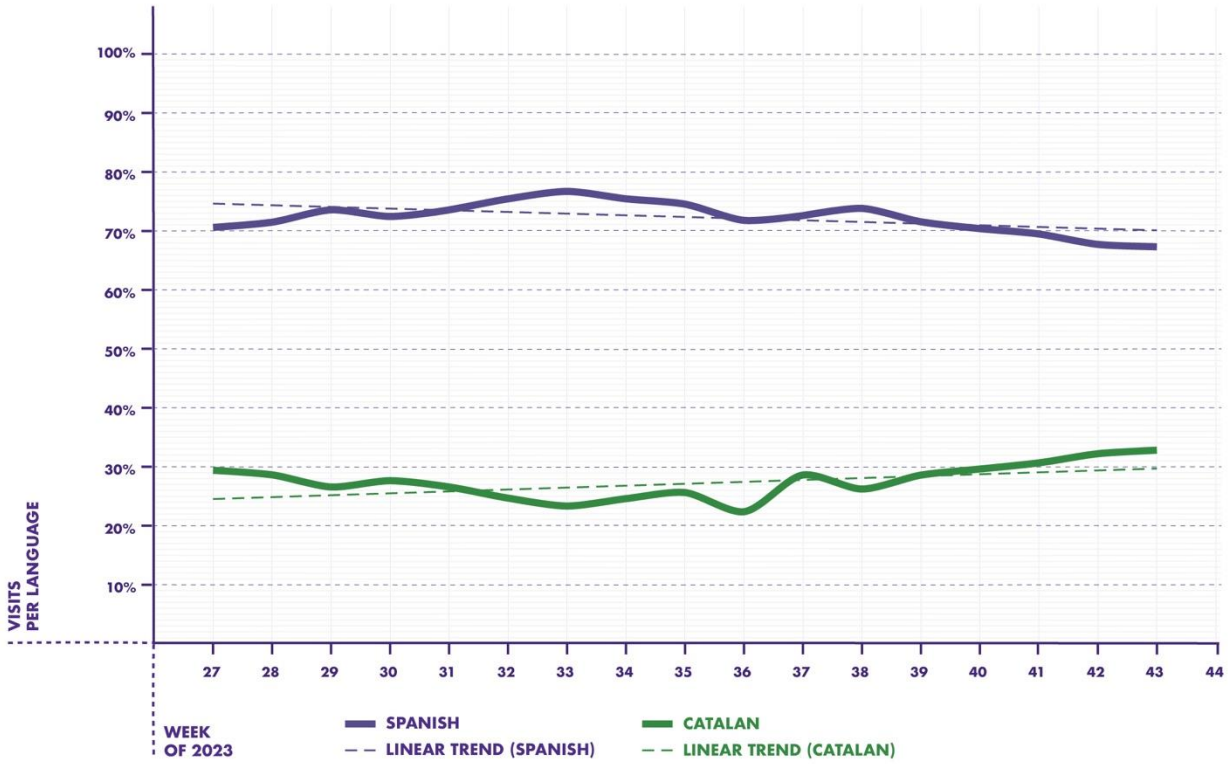
DESCRIPTION

Catalan retail material sales company which has physical stores and online stores in Catalan and Spanish.

Correlation CA – ES -1
Correlation CA – EN

Does it use "hreflang"?
Yes, no errors. Possible improvement: adding x-default.

Full series
(X-axis: week; Y-axis: percentage of traffic from Google for each language)



Collaborator 4 only works in Catalan and Spanish on its website, which means that the correlation between the two languages is the complete reverse.

In this second case, the minimum value of traffic to the Catalan versions was reached in week 33 with a value slightly over 23% (23.15%). This is a loss of over 5% in comparison

with the average traffic of the first 4 weeks, causing the Spanish version to reach its peak of almost 77% (76.85%) of visits.

As soon as Google started to apply improvements, both languages went through an initial rebound with the Catalan version reaching 28%, i.e. the same situation as at the start of the series, despite this not being consolidated. Two weeks later, however, after the end of week 38, there was a second rebound which consolidated an increase in traffic in Catalan that exceeded 33% (33.36%) at the end of the series.

On analyzing the full series, based on the average from the 4 initial weeks and from the 4 final weeks, the Catalan version consolidated an increase in traffic that exceeded 3%, which is the same value as that lost by the Spanish version.

COLLABORATOR 7

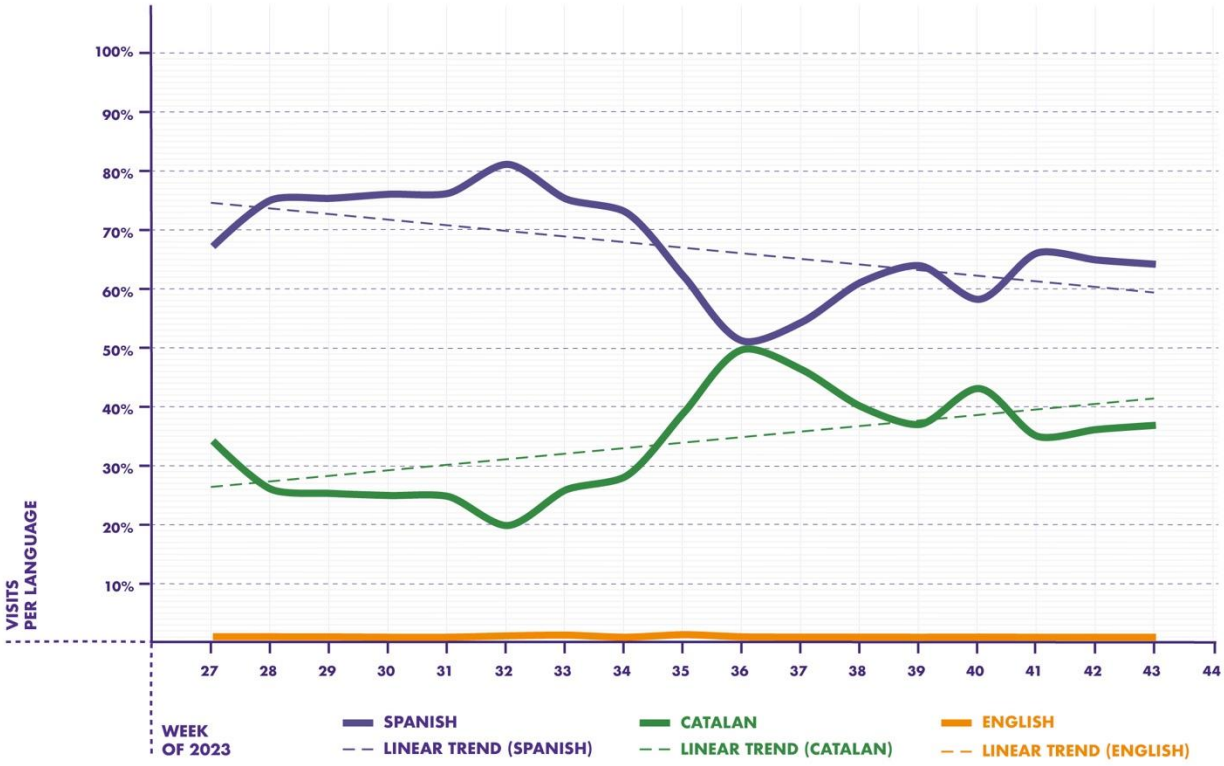
DESCRIPTION

Generalist digital press following current affairs with non-symmetrical editions in Catalan, Spanish and English.

Correlation CA – ES -0.99
Correlation CA – EN 0.01

Does it use "hreflang"?
Yes, it formats correctly.

Full series
(X-axis: week; Y-axis: percentage of traffic from Google for each language)



The case of the websites of collaborator 7 can be considered the clearest paradigm of the changes arising from the evolution experienced by the Google algorithm: major fluctuations in traffic, with extremely notable rebounds that result in a significant change in language-based trends in a matter of days.

In this case, English is left out because, at its maximum levels, it accounted for 0.3% of the total website traffic. Furthermore, significant differences in this traffic cannot be detected, since it began below 0.1% and ended up right at this value.

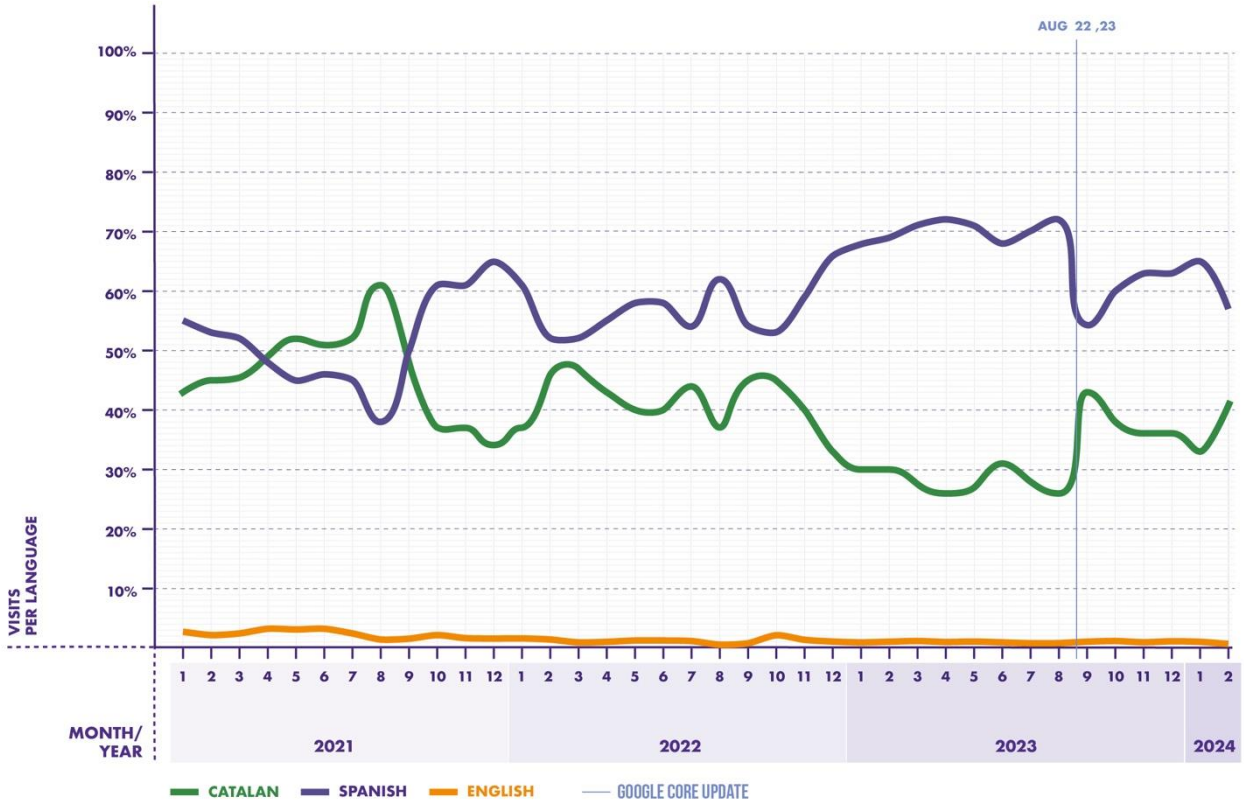
Focusing on the behavior of Catalan, it started the series (average of the first 4 weeks) with a weight of 27% of traffic, whereas the remaining 73% was traffic to the Spanish version. As time went by, the Catalan version fell to just over 19% (19.09%) in week 32, coinciding with the maximum level of the Spanish version which rose to almost 81%.

From week 33 and, above all, week 34, when Google made the changes to its algorithm, there was a sudden change in trend with an extremely strong rebound making the Catalan version rise to almost the same level as the Spanish (49.15% in comparison with 50.75%). This did not continue, however, and it initially fell until week 39 and, after a gentle rebound, fell again until week 41, after which time it began to stabilize.

On reviewing the series as a whole and comparing the average of the first 4 weeks with that of the last 4, the Catalan version rose by almost 10%, which is exclusively from the Spanish. It must be noted that the first sudden drop between weeks 27 and 28 has significant weight on this comparison.

EVOLUTION OF CATALAN IN THE COLLABORATORS AS A WHOLE

On evaluating the medium-term trend of traffic to the Catalan version of the contributing websites, the data available from a long series of at least 23 months have been added. These data have been weighted, respecting the weights provided by each collaborator to reach a **volume in excess of 400 million sessions**. The resulting diagram is shown below:



Initially **noteworthy are the fluctuations** in the behavior of traffic in Catalan and Spanish, with sudden increases and decreases, often **coinciding** with the different **algorithm updates** by Google[6]

In terms of the evolution of traffic to the Catalan version, this experienced an initial growth stage to reach its maximum value of the series, in excess of 61%, in August 2021. From then on, it fell heavily to reach an initial minimum of around 34% in December 2021, having lost almost 9% since the start of the series. It then recovered to its initial levels, yet

[6] See Annex 1 for the chronology of updates of the Google algorithm over the period considered.

maintaining notable variations until October 2022 when it started to free-fall until the summer of 2023.

The Catalan series then reached its minimum values at slightly under 27%, which is when the distance between Catalan and Spanish was at its greatest, reaching 46%. At that point, the Catalan version had lost over 16% of visitors in comparison with the first half of 2022, or more than 34% in comparison with its peak, which is a loss of 1 out of every 3 visitors.

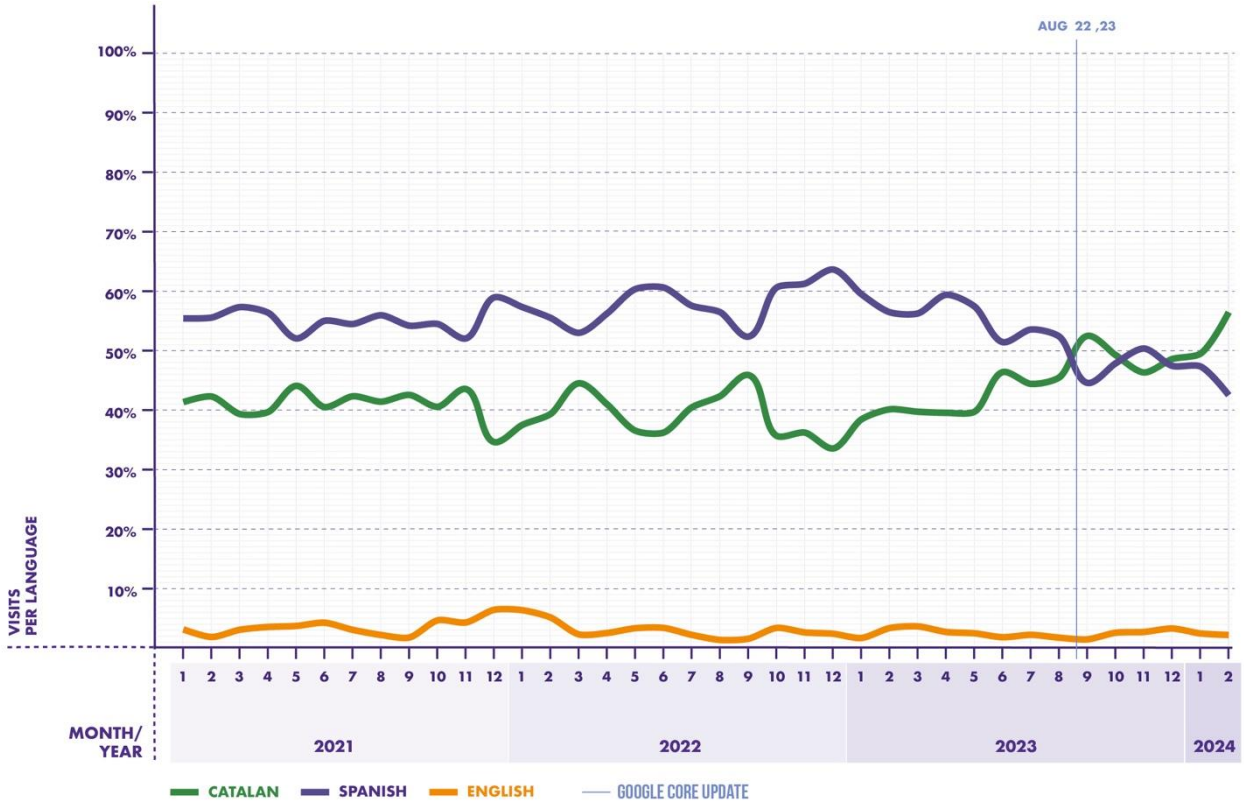
Between September and November 2023, there was a sudden rebound in traffic in Catalan, which ended with a lower decline that did not alter an upwards change in trend. **This change was consolidated between December 2023 and February 2024**. **The coming months must be monitored** to specify whether this is consolidated or whether it is merely a rebound, like in the past.

In terms of the other languages, the behavior of the Spanish version is almost symmetrical to that of the Catalan one, reaching its minimum of 38% when the Catalan version reached its peak in August 2021. Likewise, the Spanish version reached its peak in August 2023 at almost 43%, whereas the Catalan one was at minimum levels. The English version had a general downward tendency over the series, starting at around 2% of traffic, but ending up at around 1%. Its minimums and maximums do not coincide with any special event from the other series.

EVOLUTION OF CATALAN IN THE COLLABORATORS AS A WHOLE, RULING OUT THE MEDIA

On observing the behavior of the collaborators separately, it is seen that the media shows specific trends related to current affairs. This means that, beyond the impact of the changes by Google, some rebounds are linked to the publishing of news that might alter the results. The same series are therefore analyzed but ruling out the traffic from these types of collaborators.

In this case, the series is analyzed based on **approximately 100 million sessions** and is illustrated in the following diagram:



Its appearance is one of a set of series with narrower fluctuations and with greater presence—although to an extremely minor extent—of English. The most noteworthy feature is, however, **that Catalan reached its peak at the end, exceeding 56% in February 2024**.

The overall behavior of traffic toward the Catalan versions had an initial period of almost complete stability, which coincided with the first 11 months of 2021 and remained at around 41%. From then on, it entered a variable period that continued until the end of the series. This period began with a gentle downward trend but reached a peak in August

2022 when traffic in Catalan exceeded 46%. After this, **there was a sudden drop** which led to the minimum value of the series, leaving Catalan below 34%, which, once again, coincided with the time of maximum presence of Spanish at around 64%. All this took place in December 2022.

During the last part of the series, **Catalan changed trend and started on a decidedly upward path**: supported by certain changes to the Google algorithms, it exceeded traffic to the Spanish versions in September 2023 and reached the first peak of the series, exceeding 53% and coinciding with a decline in the penetration of Spanish, which stood at below 43%.

This situation is initially symbolic: the following months see a drop in Catalan to around 46% and Spanish exceeding it again, although this trend is interrupted in December. At present, **Catalan has been above Spanish for 3 consecutive months**, coinciding with the time the former reached its peak, over 56%, and the latter its absolute minimum, below 43%.

In general, from the start to the end of the series, Catalan has gained 15% of traffic (+37% in relative terms), which is almost entirely from Spanish. English shows an almost neutral situation, with an average of 3% of traffic. **In relation to the most severe period of the visibility crisis (34% of traffic), Catalan has recovered 22% in its presence** (+65% in relative terms) after the core update by Google in late August 2023.

As indicated in the previous section, monitoring of the evolution of this situation must continue, since Catalan remained the second language in terms of traffic for 35 of the 38 months reviewed. This means that, to be able to confirm that this is not a rebound, the current situation must stabilize so that it can be validated more extensively.

CONCLUSIONS

The main conclusion of this new study is that **the core update that Google applied in late August 2023 reversed the downward trend in the visibility of content in Catalan**, which is now showing signs of recovery with a clear turning point after said date.

Specifically, as of late February 2024, an increase of 15% is seen in comparison with the initial situation (in 2021) and of 22% in comparison with the worst times of the visibility crisis (late 2022 - mid 2023), which stood at 18% at the initial change in trend.

In the individual analysis of the most relevant websites, variations in the times are observed, which in some cases are considerable. There may be different reasons for this: from specific core updates by Google[7] to changes in the structure and content of the websites themselves, as well as changes in the SEO strategy.

Despite this, the trend lines show a **general tendency toward the recovery of Catalan.** What is more, just as we did when Google applied its core update, we would like to thank the company for its involvement, but would invite it to continue investigating, since we believe there is still **room for improvement in terms of the range of results.** This is specially true in view of the approach toward the Search Generative Experience, which will use generative artificial intelligence (GenAI) and will make it difficult to establish correlations between the core updates and the results.

In terms of user demand, the Alliance proposes different monitoring activities on the presence of Catalan, combined with **promotional campaigns for the linguistic normalization of the Catalan language in devices**, which will increase the potential market of recipients of search results in Catalan.

[7] See Annex 1.

NEXT STEPS

This report will be sent to the Google Engineering department to validate the positive effects of its involvement.

It will also be sent to the members of the Alliance, to the Governments of Catalan-speaking territories, and to the contributing organizations of the study, and is published on the Alliance website available to the public. Once the new European Parliament has been established after the elections of June 6 to 9, 2024, it will also be sent to the Minority Intergroup.

The study will be updated in 12 months to verify whether the improvement trend continues. Where a new decline in the visibility of Catalan is evident, the possibility of moving the new study forward will be considered.

Alongside this, **the .cat Foundation has started to deploy the Xarxa de Monitoratge del Català Digital** (Catalan Digital Monitoring Network, XMCD)[8], formed by sensors distributed around all Catalan-speaking territories—and in other places around the world, to a lesser extent—which regularly conduct Internet searches and inform the server of the .cat Foundation of the presence of content in Catalan among the initial search results. A dashboard shows how the visibility of Catalan evolves over time in terms of Internet users and detects any declines.

Entities linked to different languages around the world, which are also potentially vulnerable to core updates, have also shown an interest in deploying their own monitoring networks based on the technology of the XMCD. The .cat Foundation is studying the options of contributing to these deployments.

Alongside this, in view of **the growing use of AI chatbots** as an interface for access to digital information, work will begin on **developing a methodology** to calibrate the visibility of content in Catalan within this new environment.

Finally, in light of the evidence that many Catalan-speaking users do not have their Internet browser environment correctly configured in Catalan, it is believed that **increasing the proportion of devices with the linguistic normalization of the Catalan language** is explicitly the most effective way of informing Internet browsers of the desire to receive results in Catalan. Alliance members therefore propose the short-term promotion of the linguistic normalization of the Catalan language in devices—stationary and mobile, for personal and corporate use—through a combined information campaign that promotes the

[8] https://xmcd.fundacio.cat

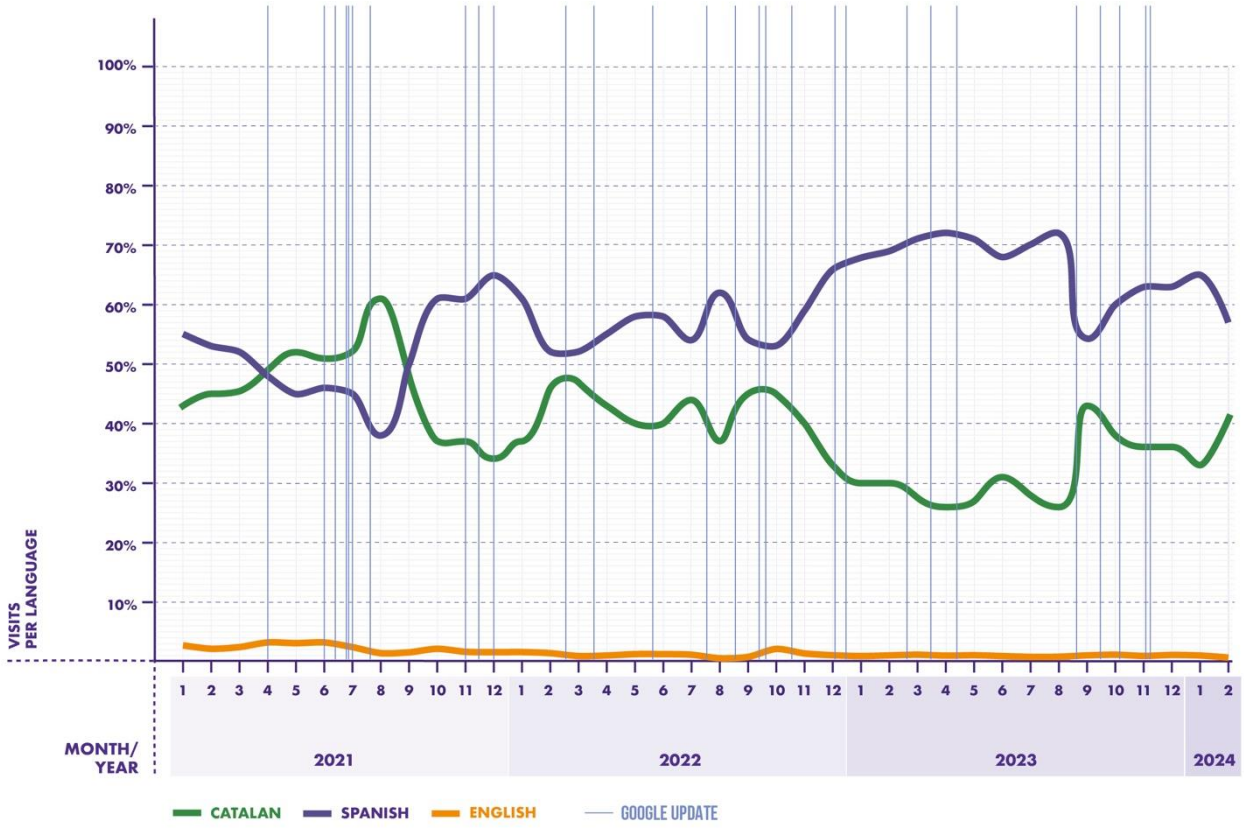
use of existing configuration tools, such as the "Catalanitzador" by Softcatalà—a member of the Alliance—or of newly created tutorials.

ANNEX 1. GOOGLE UPDATES, A MOVING TARGET

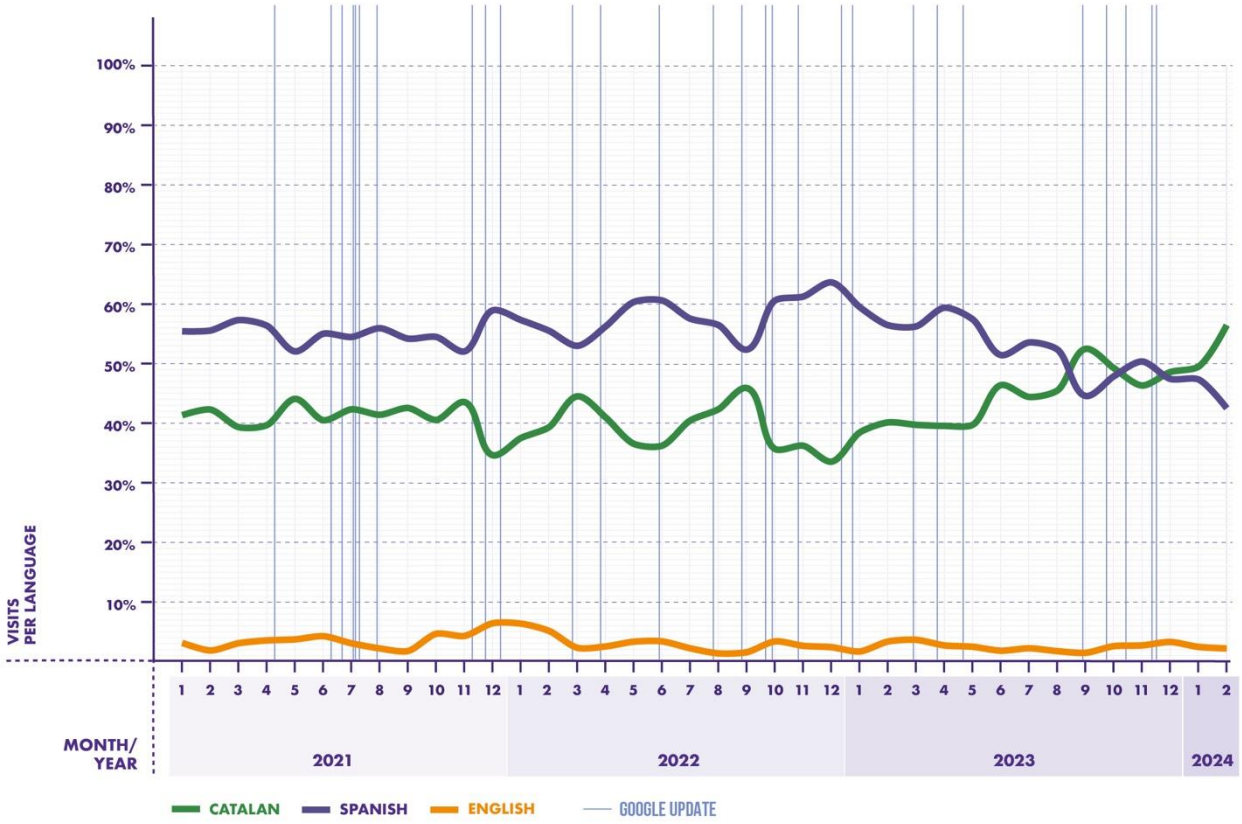
During our studies, we observed the effects of the successive core updates by Google. Beyond the main **core updates**, the company is constantly making changes to adapt the behavior of search results, sometimes in response to practices it believes are attempting to manipulate them.

For informative purposes, we have superimposed the chronology of core updates that Google made between January 2021 and November 2023 over our graphs showing cumulative traffic received for each language: each vertical line corresponds to an update.

The first, for all contributing websites, with 400 million sessions:



The second, excluding the media, with 100 million sessions:



The graphs show how **some** of the updates by Google **directly affect** language-based visibility, whereas **others have no effect**.

In any case, the monitoring of these updates by Google must form part of future studies.

CREDITS

- Report Authors: Albert Cuesta (albertcuesta@fundacio.cat), Pep Masoliver (jmasoliver@fundacio.cat)
- Technical Management, Data Processing: Pep Masoliver
- Data Acquisition: Jaume Medina (jmedina@fundacio.cat), Griselda Casadellà (gcasadella@fundacio.cat)
- Report Edition: Albert Cuesta, Griselda Casadellà
- Graphics and layout: Amets Díaz
- Communication: Meritxell Alavedra (malavedra@fundacio.cat), Maite Bassa (mbassa@fundacio.cat)

MEMBERS OF THE ALIANÇA PER LA PRESÈNCIA DIGITAL DEL CATALÀ

- Acció Cultural del País Valencià: Anna Oliver, Vicent Ferrer
- Amical Wikimedia: Àlex Hinojo, Núria Ribas
- Fundació .cat: Genís Roca, Roger Serra, Gerard Vélez
- Institut d'Estudis Catalans: Àngel Messeguer, Francesc Salvador
- Institut Ramon Llull: Àlex Hinojo, Pere Almeda
- Obra Cultural Balear: Llorenç Garcia
- Òmnium Cultural: Iker de Luz
- Plataforma per la Llengua: Marc Biosca
- Softcatalà: Joan Montané
- WICCAC: Joan Soler

COLLABORATORS

The Aliança per la Presència Digital del Català and the report authors wish to thank each and every one of the organizations that answered our call to provide traffic data from their websites in order to be processed for this report.

Some of these entities have expressly agreed to be mentioned here:

Ajuntament de Barcelona

Amical Wikimedia

Eurecat, Centre Tecnològic de Catalunya

Generalitat de Catalunya

Institut Ramon Llull

Meteocat, Servei Meteorològic de Catalunya

Òmnium Cultural

Universitat Pompeu Fabra

Universitat de Barcelona

Universitat de Girona

Universitat Oberta de Catalunya

Fundació Mobile World Capital Barcelona

Futbol Club Barcelona

CONFIDENTIALITY

We do not publish any other name here in order to preserve their privacy, as indicated in the Non-Disclosure Agreement signed with each of the contributing entities. However, such information could be disclosed to one or more web search providers, should they require it to analyze a specific case more deeply. This disclosure will only be made on a case-by-case basis and only after the Contributor has agreed to it, as it is also covered by the NDA.

Aliança per la presència digital del català

Plaça Nova, 7 ,5a planta,
08002 Barcelona
354 750 936

info@aliançadigital.cat

ACCIÓ CULTURAL DEL PAÍS VALENCIÀ · ARTICLE WIKIPEDIA · fundació .cat · INSTITUT D'ESTUDIS CATALANS MCMVII · institut ramon llull · OBRA CULTURAL BALEAR · ÒMNIUM LLENGUA CULTURA PAÍS · PLATAFORMA PER LA LLENGUA · SOFTCATALÀ · WIKIMEDIA CCAC