# Surveying Perspectives on Language Technology and the Perceived Needs of Moroccan Arabic and Tamazight Speakers

**Avery C. Kanel***
Saarland University
Campus 66123 Saarbrücken, Germany
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3 66123 Saarbrücken, Germany
[avka00001@stud.uni-saarland.de]

**Christian Schuler***
Saarland University
Campus 66123 Saarbrücken, Germany
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3 66123 Saarbrücken, Germany
[christianschuler8989@gmail.com]

**Yusser Al Ghussin**
Saarland University
Campus 66123 Saarbrücken, Germany
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3 66123 Saarbrücken, Germany
[yusser.al_ghussin@dfki.de]

**Yassine Chaouri**
National Institute of Posts and Telecommunications
avenue Allal Al Fassi, Madinat Al Irfane, Rabat, Morocco
[yassinechaouri0@gmail.com]

**Bader Ait Ammi**
Université Sultan Moulay Slimane
B.P. 591, Av. Ibn Khladoun, 23000 Beni Mellal, Morocco
[badraitammi393@gmail.com]

**Simon Ostermann**
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3 66123 Saarbrücken, Germany
[simon.ostermann@dfki.de]

**Josef van Genabith**
Saarland University
Campus 66123 Saarbrücken, Germany
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3 66123 Saarbrücken, Germany
[Josef.van_Genabith@dfki.de]

**Abstract**

Morocco is a country home to around 34 million speakers of Moroccan Arabic (Darija) and nine million speakers of the Moroccan Tamazight varieties (Central Atlas Tamazight, Tarifit, and Tachelhit). Language technologies are increasingly being applied to non-standardized languages and dialects. However, outside of Europe, very few countries have conducted NLP community needs surveys for unofficial and minority languages. Darija and Tamazight fall, respectively, into those categories. We survey speakers of these languages (N=29) and ask both about their attitudes towards emerging language technologies and which ones they would find useful. We found the majority of respondents were Darija-native speakers that spoke English, who shared a positive view of emerging language technologies, especially chatbots. We also noted the community was split on the topic of spell and grammar checkers and standardized spelling for Darija. While the survey captured Tamazight speakers, all respondents reported Darija as the language they felt closest to.

## Introduction

Morocco is home to a dense and layered ecology of language varieties, Darija and Tamazight being two that are under pressure and changing from diglossia (Fathi, 2024) and disappearance of dialectical diversity (Bao, 2026; Brandea, 2025). Recent scholarship describes Tamazight in Morocco as moving from marginalization toward institutionalization (Idhssaine, 2022), while also noting that implementation remains uneven across domains such as media and public communication (Belhiah et al., 2020).

Darija, spoken by nearly all of Morocco's 37 million inhabitants as a first or second language (Heath, 2020), appears robust by speaker count. Yet its rich regional diversity (phonological, lexical, and contact-driven) (Aguadé, 2008; Boumans, 2005), is being leveled by urbanization-driven koineization (Falchetta and Guerrero, 2023; Hachimi, 2012), the process by which a new simplified dialect emerges in a city where speakers of different, mutually intelligble dialects mix, while written and prestigious domains remain occupied by Standard Arabic (MSA) and French in a layered diglossia with pervasive code-switching (Skiredj et al., 2024; Hermas, 2025). Erosion thus operates at two levels at once, with the endangerment of the Tamazight languages, and the quieter disappearance of dialectal diversity within Darija itself, which language resources routinely flatten into a single "Moroccan Arabic" label (Heath, 2013).

*Contributed equally.

We present a community-centered survey inviting speakers of Darija and Tamazight varieties to share how they relate to their language, what they want and do not want from language technology and research. Following critiques of top-down design, the survey asks the following without presupposing that standardization or increased digitization are self-evidently desirable.

1. **Language Use:** Which languages the respondents know and in what settings they use them.
2. **Language Beliefs:** Respondent attitudes on variation in and standardization of their languages.
3. **Language Technologies:** Priorities of respondents for future language technologies and their disposition towards the field in general.

Findings will be reported back to the participating communities via the same channels used to advertise this survey to begin with, and will directly shape what we build and study next.

## Related Work

Language technology has largely passed non-standardized and minority communities by. Joshi et al. (2020) show that the vast majority of the world's languages, Tamazight varieties among them, sit in the lowest resource classes of the NLP world and it has been prognosed that most of these languages will soon perish (Kornai, 2013). At the same time, a growing body of work argues that dialect-aware language technology is both needed and increasingly feasible (Joshi et al., 2025). Yet feasibility is not the whole question. When technology does reach non-standardized varieties, it is too often built top-down, treating them as deficient versions of high-resource languages, something Bird (2022) critiques as "poverty-conscious" design and importing assumptions about standardization, writing, and desirable applications that speaker communities may not share. Community surveys in other contexts have repeatedly overturned expectations.

The DLDP survey of four European minority languages found a strong desire for digital presence, held back less by disinterest than by technological, linguistic, and psychological barriers (Soria et al., 2018). Millour (2019)'s survey of Alsatian debunked the stereotype of an aging speaker base and found that informal cultural hubs reached speakers far better than institutional channels. Blaschke et al. (2024) found that German dialect speakers welcome technology that understands their dialect but resist technology that writes it, with 65% opposing a standardized orthography. Finally, Liu et al. (2022) caution that "low-resource" is not a monolith but rather that community priorities often diverge sharply from researcher benchmarks. To our knowledge, no comparable community survey has yet engaged speakers across Morocco's linguistic landscape.

## Community Survey

The survey is divided into four sections, one as a screener to filter out non-Darija and non-Tamazight speaking respondents. The other three dealt with the respondent's use of their language and personal information, their attitude towards their language, and priorities for language technologies respectively.

For the localization of the survey's contents, we started with Morocco's high-resource languages (English, French, Spanish, and Standard Arabic). In collaboration with our contacts in local language communities, we translated the survey's contents into Darija and Tamazight which allows monolingual speakers to participate as well. Delivering a survey about dialectal technology in the dialect itself signals that these varieties are fully capable of digital life (Soria et al., 2018). To better reflect the reality of language communities and their geographic distributions, we do not provide rigid language labels for participants to select from, but rather ask participants to write down the name of their language and location where it is spoken.

Speakers of marginalized varieties often shift to dominant languages not from lack of attachment, but because prestige and functional range outweigh mother tongue preference in daily life; and in comparable communities, a majority believe technological support could make their language more attractive to younger generations (Blaschke et al., 2024). Hence, dialect-aware technology, built with communities, may help restore the everyday utility and prestige whose absence erodes generational transmission. Whether communities themselves see technology in this role is, fittingly, one of the questions the survey asks rather than assumes.

## Results

We present the survey design, lessons from the localization process, and first results from our data collection efforts. It should be noted that participants could skip all but the first two questions in the survey. Results presented below are in percentages, we report responses in Figure 2 and Figure 3.

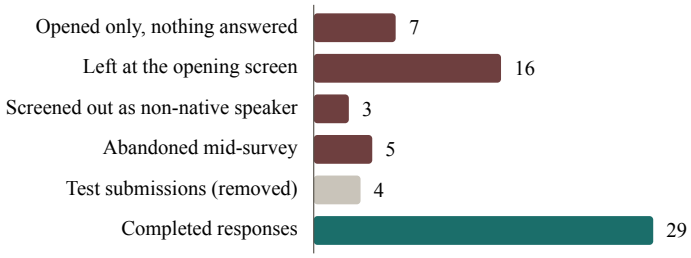



Figure 1: Recorded participations with sessions lost before completion in brown, while teal shows the 29 completed responses analyzed here. Median completion took 9.2 minutes across all sessions.

As of 14 September 2026, the server held 64 sessions (Figure 1): 15 visitors who left at the opening screen, four opened without answering, five abandoned mid-survey, three of them only at the technology block, after completing all demographic and attitude questions, and 29 fully completed responses, with a median completion time of 9.2 minutes. Social media was the largest recruitment channel (12 of 29), ahead of acquaintances (7) and researcher contacts (5), consistent with Millour (2019)'s observation that informal channels reach speakers better than institutional ones. From all 29 respondents who answered the daily language question, 20 named Darija and 6 Tamazight or Tachelhit. Although all 29 respondents listed Darija as the language they felt closest to. The preference for Darija is explored more in subsection . Thus conclusions drawn in our Results will describe more the state and wants of speakers of Darija rather than the Tamazight languages. Table 1 and Table 2 present response statistics from the survey.

| | $n = 29$ | % |
|---|---|---|
| **Region (Q 3)** | | |
| other regions (1 each) | | |
| Casablanca | 8 | 28 % |
| no answer | 8 | 28 % |
| Morocco (unspec.) | 2 | 7 % |
| Oriental region | 2 | 7 % |
| | 9 | 31 % |
| **Daily languages (Q 4)**† | | |
| English | 27 | 93 % |
| French | 24 | 83 % |
| Darija | 20 | 69 % |
| Modern Standard Arabic | 6 | 21 % |
| Tachelhit / Tamazight | 6 | 21 % |
| Spanish | 4 | 14 % |
| **Writes their language (Q 9)** | | |
| yes | 23 | 79 % |
| no | 4 | 14 % |
| **Script used for it (Q 29)** | | |
| Latin | 19 | 66 % |
| Arabic | 4 | 14 % |

Table 1: Linguistic profile of the completed responses, as counts and rounded percentages of $n$. Regions are hand-coded from free-text answers. "No answer" includes fields optionally hidden based on previous user responses in the survey.

**Unpopularity of Forms of MT and Spell and Grammar Checkers** Interestingly, out of all technologies respondents were surveyed on, machine translation (MT) from Standard Arabic into Darija was the least popular, with over 40% of respondents indicating that the technology was *'somewhat not useful'* or *'not useful'*. What is striking about this result is that machine translation from Darija into Standard Arabic is perceived much more positively, with only 16% of respondents indicating the technology as *'somewhat not useful'* or *'not useful'*. Another unpopular technology is spell and grammar checkers, with around 40% of respondents saying the technology is *'somewhat not useful'* or *'not useful'*.

| | $n = 29$ | % |
|---|---|---|
| **Age (Q 12)** | | |
| 20–29 | 16 | 55 % |
| ≤19 | 9 | 31 % |
| 30–39 | 2 | 7 % |
| **Education (Q 5)** | | |
| Licence | 9 | 31 % |
| Bac | 8 | 28 % |
| Master+ | 6 | 21 % |
| other / not disclosed | 6 | 21 % |
| **Heard about the survey (Q 28)** | | |
| Social media | 12 | 41 % |
| Acquaintances | 7 | 24 % |
| Researchers | 5 | 17 % |

Table 2: Demographics and recruitment channel of the participants who completed the survey.

**Popularity of Chatbots** The most popular technology according to respondents was the ability for chatbots to both understand and respond in Darija, with over 80% responding in favor of it.

**Speech-to-Text and Text-to-Speech** Speech-to-text drew a notable share of neutral responses and positive responses but no negative ones (64% were positive about Darija-to-MSA, 75% for Darija-to-Darija). Similarly 77% of respondents were postive about text-to-speech technology.

**Receptiveness to Language Technologies** All respondents showed a positive attitude towards language technology, with 100% indicating that if they found a technology useful, they would use it. 62% said that more language technology would make the respondent use their Darija more. 76% said more language technology could make their language more appealing to youth. There is likely a dependent relationship between these two questions, as only 7% of the respondents were above the age of 29, while 31% were younger than 19.

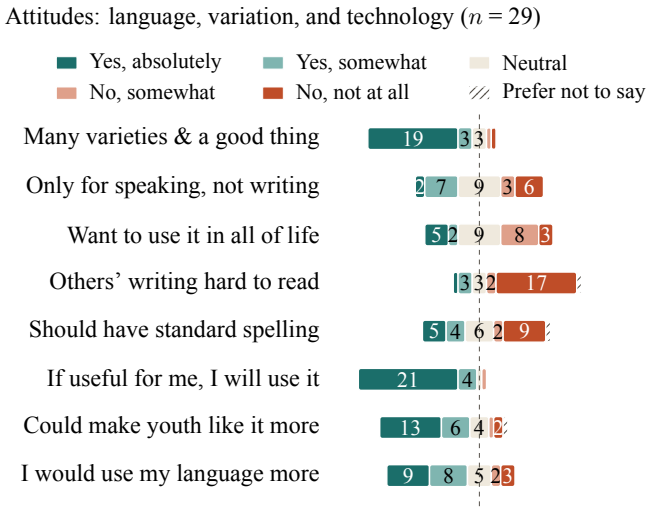



Figure 2: Attitudes toward language variation, writing, and standardization (Q 14 and Q 26).

**Attitudes Towards Darija** Respondents did not find other's writing in Darija hard to read and there were mixed reactions to the sentiment that Darija is *"Only for speaking, not writing"*, with 33% in favor of, 33% neutral on, and 33% opposed to the idea. Having *"standard spelling"* was similarly divisive, with 34% in favor of, 23% neutral on, and 42% opposed to it. Regarding respondents' attitudes towards Darija, 80% said that the varieties of Darija were a positive thing in their language. These numbers suggest that speakers embrace and value the rich dialectal variation of their language, which would run counter to the notion of introducing a standardized language, while simultaneously expressing the wish for

having a standardized spelling which should then happen on a more local level of regional dialects rather than country-wide.

**Sub-Population of Respondents** 93% of respondents reported daily use of English and 83% reported daily use of French. This is important to mark, as it indicates that the respondents to the survey are a sub-population of Morocco, whose national average English reading and writing population is 20.5% of the literate population (Haut-Commissariat au Plan, 2024).

**Language Technology Popularity Ranking** Orthographic fidelity shows an interesting asymmetry in that for machine translation output, respondents lean slightly toward not caring which way of writing Darija is used (ten vs. eight who want their own), but for chatbot replies a majority wants fidelity (fourteen of twenty-seven want answers matching their own writing, only four do not care). 77% of respondents know of no language technology for Darija; those who do point to general-purpose AI systems, e.g. *"kandn chatGPT kay3rf yktb b darija"* ("I think ChatGPT knows how to write in Darija"). In an exploratory analysis we find that respondent's identity attachment to their language predicts support for language technologies, which means that the more central Darija is to who a respondent is, the more technology they want for it. While in contrast, attitudes toward standardization matter only for technologies whose output is written in Darija.

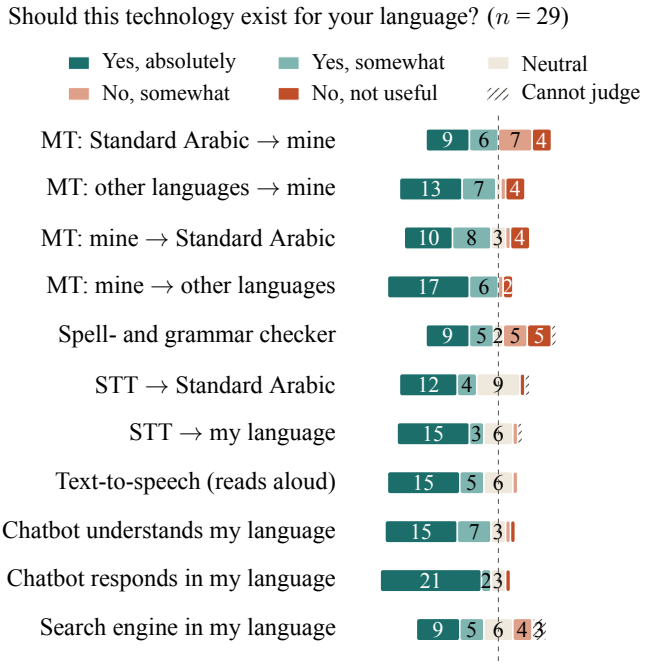



Figure 3: Attitudes toward support for individual language technologies (Q 16–Q 23).

**Tamazight and Highly-Multilingual Speakers** Finally, two exploratory subgroups indicate possible heterogeneity within the sample that is worth re-testing as the sample size grows. The five respondents who use a Tamazight variety daily name Darija as their closest language, showing that although the survey reached Tamazight speakers, it did not reach communities where Tamazight was the primary language. The association between identity attachment to Darija and support for norm-imposing technology on the language has a different trend for Tamazight speakers than the remaining population. The more the language matters to Tamazight speakers, the less they want standardization, spell-checking, and machine translation, contrary to other respondents. Among the thirteen respondents with more than three daily languages, the belief that technology raises the language's appeal to youth is popular regardless of their other attitudes. With subgroups this small and partially overlapping we can only say that these are candidate interactions and not finding. They define what the next collection wave, and especially the hoped-for Tamazight-first respondents, will let us test.

## Conclusion & Outlook

### Discussion

In our survey of Darija and Tamazight speakers we found that most respondents viewed language technologies favorably, although two notable exceptions were translation from Standard Arabic into dialectical as well as spell and grammar checkers. The disinterest in Standard Arabic to dialectal machine translation is likely due to the average, educated Moroccans passive proficiency in Standard Arabic but weaker than native linguistic proficiency (Ech-Charfi and Bouaddi, 2023). The disinterest in spell and grammar checkers is likely due to Darija's unstandardized orthography and use in colloquial rather than professional settings. Recent work has begun building conversational models specifically for Darija (Shang et al., 2025), reflecting growing interest in dialectal chat interfaces and seems to align with the positive responses from respondents in language technologies related to chatbots. Speech-to-text technologies were about as popular as text-to-speech technologies which runs counter to what previous dialect and spoken language studies found when working with groups from other language communities (Blaschke et al., 2024; Chrupała, 2023).

### Ethical Concerns

The collection of data was anonymous, voluntary and not associated with a reward of any kind. Similar to the work of Millour (2019), this survey is also based on voluntary crowd-sourcing for ethical reasons and ensuring quality (Fort et al., 2011), as well as for practical reasons of access to the participants (Callison-Burch and Dredze, 2010).

### Limitations

While the survey had both Darija and Tamazight speaking respondents, all respondents chose Darija as the language they were closest to. This can be partially attributed to the majority of respondents, who identified their location, locating themselves in Casablanca-Settat, a place where the percentage of mother-tongue Tamazight speakers is only 3.6% versus the national average of 18.9% (Haut-Commissariat au Plan, 2024). The survey was largely spread through online platforms, so the high population of Casablanca-Settat coupled with it's highest rate of internet use in the nation may explain the preponderance of respondents from that specific region. In addition, prior work noted Darija's role as a marker of "Moroccanness" that may influence respondents to feel closer with Darija rather than Tamazight (Caubet, 2017). In order to capture Tamazight-speaking respondent's opinions we argue for increased emphasis on non-digital means (e.g. exporting the survey as pdf-files and using print-outs) of awareness and targeting regions of high Amazigh population.

## Acknowledgments

We thank all participants from the language community. Christian Schuler gratefully acknowledges funding from GraduSaar, Saarland University.

## References

Aguadé, J. (2008). Morocco [dialectological survey]. In Versteegh, K., Eid, M., Elgibali, A., Woidich, M., and Zaborski, A., editors, *Encyclopedia of Arabic Language and Linguistics*, volume 3, pages 287–297. Brill, Leiden–Boston.

Bao, K. (2026). *Language policy and planning of Amazigh languages in Morocco: a study of the language ideology of the Royal Institute of Amazigh Culture (IRCAM)*. PhD thesis, Universiteit Leiden, Leiden.

Belhiah, H., Majdoubi, M., and Safwate, M. (2020). Language revitalization through the media: A case study of amazigh in morocco. *International Journal of the Sociology of Language*, 2020(266):121–141.

Bird, S. (2022). Local languages, third spaces, and other high-resource scenarios. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Blaschke, V., Purschke, C., Schütze, H., and Plank, B. (2024). What do dialect speakers want? a survey of attitudes towards language technology for german dialects.

Boumans, L. (2005). Lexical variation in Moroccan Arabic: The genitive exponent in experimental data from four cities. *EDNA, Estudios de dialectología norteafricana y andalusí*, 9:125–137.

Brandea, B. (2025). A history of moroccan arabic studies. *Romano-Arabica*, 24(1):79–95.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In Callison-Burch, C. and Dredze, M., editors, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles. Association for Computational Linguistics.

Caubet, D. (2017). Darija and the construction of "Moroccanness". In Bassiouney, R., editor, *Identity and Dialect Performance: A Study of Communities and Dialects*, chapter 6. Routledge, London.

Chrupała, G. (2023). Putting natural in natural language processing. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827, Toronto, Canada. Association for Computational Linguistics.

Ech-Charfi, A. and Bouaddi, Y. (2023). The receptive vocabulary size of moroccan learners of modern standard arabic. *Journal of Applied Language and Culture Studies*, 6(3):1–19.

Falchetta, J. and Guerrero, J. (2023). Koineization trends in spoken arabic across three moroccan towns. *Kervan – International Journal of Afro-Asiatic Studies*, 27(1).

Fathi, S. (2024). Dual diglossia in morocco: A new sociolinguistic situation. *International Journal of Language, Linguistics, Literature and Culture*, 3(4):1–9.

Fort, K., Adda, G., and Cohen, K. B. (2011). Last words: Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Hachimi, A. (2012). The urban and the urbane: Identities, language ideologies, and arabic dialects in morocco. *Language in Society*, 41(3):321–341.

Haut-Commissariat au Plan (2024). Recensement général de la population et de l'habitat 2024.

Heath, J. (2013). *Jewish and Muslim Dialects of Moroccan Arabic*. Routledge, London.

Heath, J. (2020). Moroccan arabic. *Arabic and contact-induced change*, 1:213.

Hermas, A. (2025). Managing Morocco's linguistic market: A linguistic citizenship proposal. *European Journal of Language Policy*, 17(2):237–254.

Idhssaine, A. (2022). The evolution of the status and teaching of amazigh in morocco: From marginalization to institutionalization. *Journal of Language Teaching*, 2(12):1–7.

Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., and Dippold, D. (2025). Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6):149:1–149:37.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kornai, A. (2013). Digital language death. *PLOS ONE*, 8(10):e77056.

Liu, Z., Richardson, C., Hatcher, R., and Prud'hommeaux, E. (2022). Not always about you: Prioritizing community needs when developing endangered language technology. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Millour, A. (2019). Getting to know the speakers: A survey of a non-standardized language digital use. In *9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland.

Shang, G., Abdine, H., Khoubrane, Y., Mohamed, A., Abbahaddou, Y., Ennadir, S., Momayiz, I., Ren, X., Moulines, E., Nakov, P., Vazirgiannis, M., and Xing, E. (2025). Atlas-chat: Adapting large language models for low-resource moroccan arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Skiredj, A., Azhari, F., Berrada, I., and Ezzini, S. (2024). DarijaBanking: A New Resource for Overcoming Language Barriers in Banking Intent Detection for Moroccan Arabic Speakers. arXiv:2405.16482 [cs.CL].

Soria, C., Quochi, V., and Russo, I. (2018). The dldp survey on digital use and usability of eu regional and minority languages. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).