SENTimental - A Simple Multilingual Sentiment Annotation Tool

John Vidler
School of Computing
and Communications
Lancaster University
Lancaster
United Kingdom
j.vidler@lancaster.ac.uk

Paul Rayson
School of Computing
and Communications
Lancaster University
Lancaster
United Kingdom
p.rayson@lancaster.ac.uk

Dawn Knight
School of English, Communication
and Philosophy
Cardiff University
Cardiff
Wales
knightd5@cardiff.ac.uk

Abstract

Here we present SENTimental, a simple and fast web-based, mobile-friendly tool for capturing sentiment annotations from participants and citizen scientist volunteers to create training and testing data for low-resource languages. In contrast to existing tools, we focus on assigning broad values to segments of text over specific tags for tokens or spans to build datasets for training and testing LLMs.

The SENTimental interface minimises barriers to entry with a goal of maximising the time a user spends in a *flow state* whereby they are able to quickly and accurately rate each text fragment without being distracted by the complexity of the interface.

Designed from the outset to handle multilingual representations, SENTimental allows for parallel corpus data to be presented to the user and switched between instantly for immediate comparison. As such this allows for users in any loaded languages to contribute to the data gathered, building up comparable rankings in a simple structured dataset for later processing.

1 Introduction

The Small Welsh Language Model pilot (SmolLM) project was sponsored by the Welsh Government, with the main objective of addressing a significant shortfall in the available tools and models for the Welsh language. Specifically, the project aimed to tackle the lack of resources and capabilities for accurately handling sentiment analysis in Welsh. The limited availability of resources and annotated datasets for Welsh hindered the development of effective sentiment analysis tools for this language.

To bridge this gap, the SmolLM project set out to create one or more proof of concept Large Language Models (LLMs) capable of analysing Welsh texts for sentiment. This in turn required an annotated parallel corpus, from which we could use existing English sentiment analysis techniques to inform our Welsh analysis.

Ideally, the project would have utilized an existing annotated dataset to train the Welsh sentiment analyser. However, no such dataset was available in sufficient quantity or quality to support the training of robust models. This shortage of annotated data highlighted the need for a bespoke solution, where a new dataset would be created to meet the specific requirements of the project. Fortunately, Senedd Cymru (Welsh Parliament) provided a valuable resource in the form of bilingual transcripts from modern debates, which were available online as part of their Open Data[1] datasets. These transcripts offer a unique opportunity to create a large, annotated dataset for Welsh sentiment analysis, paving the way for the development of more accurate and effective sentiment analysis tools for the Welsh language.

The primary objective of the project was to facilitate the annotation of Welsh text with sentiment ratings, a crucial step in developing Natural Language Processing (NLP) models that can accurately analyse and understand the emotional tone of Welsh language content. However, a significant challenge lay in the fact that there are currently no reliable models or resources available for sentiment analysis in Welsh, hindering the development of effective NLP tools for this language.

To address this limitation, the project adopted a creative approach by leveraging an existing sentiment analyser designed for English text. By utilizing a parallel Welsh-English corpus, where the same text is available in both languages, the project harnessed the sentiment ratings generated by the English sentiment analyser to inform the training of a Welsh sentiment analysis model. This ap-

[1] https://senedd.wales/help/open-data/ - Senedd Cymru — Welsh Parliament - Open Data, accessed May 2025

Proceedings of Recent Advances in Natural Language Processing,pages 1320–1326
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-153

proach enabled the transfer of knowledge from a well-resourced language like English to a lower-resourced language like Welsh, providing a foundation for the development of a Welsh sentiment analysis model.

To further refine and improve the accuracy of the Welsh sentiment analysis model, a tool called SENTimental was developed. SENTimental is designed to collect sentiment ratings from human assessors, using the same data format as the training and testing data, to provide a more accurate and reliable source of information for the model. By incorporating human-annotated ratings, SENTimental plays a crucial role in informing the model and enhancing its performance, ultimately contributing to the development of a more effective Welsh sentiment analysis model.

2 Existing Tools

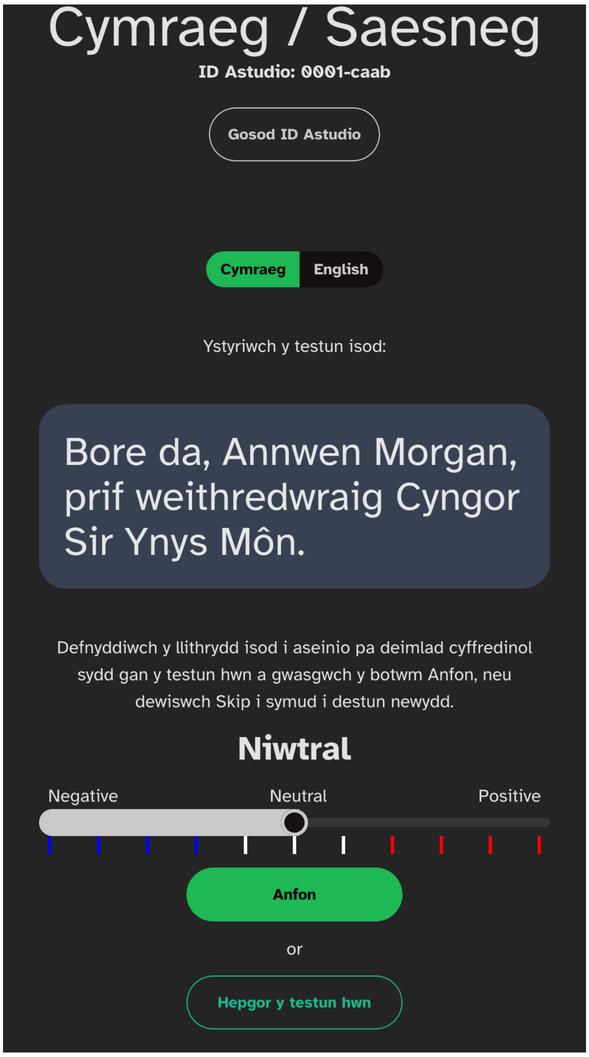
In the landscape of text annotation tools, numerous options exist that cater to general-purpose annotation needs. However, it became evident that many of these tools are either overly complex or ill-suited for the specific task of capturing sentiment ratings in the context of our research. While some tools demonstrated the capability to assess the sentiment of a passage of text, they were primarily designed for tagging tasks. These tagging tools typically require the association of discrete tokens or spans of text with specific values, which inherently complicates the process of annotating a fragment. This is most easily demonstrated by taking the number of steps required to enter a given tag; as this normally requires some form of selection-interaction (for the span) followed by a choice or generation interaction (for the tag), whereas we instead needed simple whole-fragment annotation.

Examples of such existing systems, which would be suitable for expert users but may be unsuitable for non-experts, include Docanno (Nakayama et al., 2018), INCEpTION (Klie et al., 2018), BRAT (Stenetorp et al., 2012) and Prodigy (Montani and Honnibal) all of which, while powerful, require that the user has some fairly deep understanding of the task in question, and would require not-insignificant training to have non-expert users capture the annotations correctly.

Of special note in this class of tool is Masakhane Elisa (Lin et al., 2018) which successfully handles some of the lesser-used codepoints which may be used in low-resource languages; although in this case this was not needed as Welsh uses the same Latin base set as English.

3 Design

To enhance user engagement and facilitate data entry, we prioritized the design of our user interface to be fully functional on mobile browser displays, including those found on smartphones and tablets, known as a 'Mobile First' (Mullins, 2015) design approach. This approach is intended to maximize opportunities for what we refer to as ”idle entry”, a concept that recognizes that users may not always be in a traditional data entry environment. Instead, they might be in informal settings, yet still willing to evaluate and respond to a series of prompts.



Bore da, Annwen Morgan, prif weithredwraig Cyngor Sir Ynys Môn.

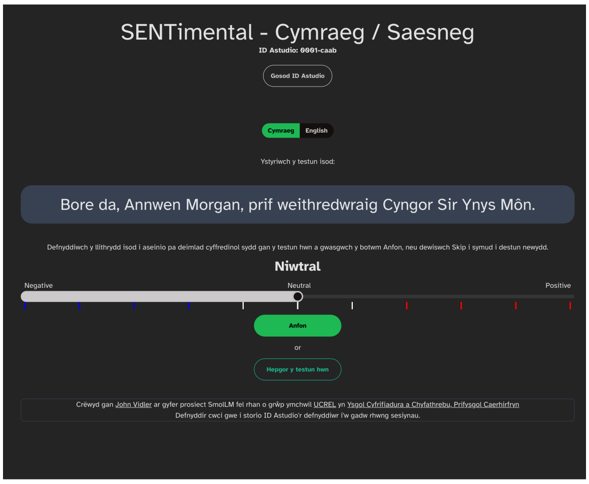
Defnyddiwch y llithrydd isod i aseinio pa deimlad cyffredinol sydd gan y testun hwn a gwasgwch y botwm Anfon, neu dewiswch Skip i symud i destun newydd.

Figure 1: The mobile UI for SENTimental, although almost any screen resolution can be used, as every element in the interface is designed to rescale to fit. All user controls are designed to be usable with fingers as well as cursors or styli.

By ensuring that the user interface is optimized for mobile devices (See 1), we provided users with the flexibility to engage on any platform. This adaptability not only makes the data entry process more comfortable but also encouraged users to participate more frequently; as low-resource languages frequently deal with smaller participant pools, we believe this to be highly relevant to maximise the data gathered with this and similar tools.

3.1 Flow State

Similarly, to maximize user engagement and productivity it is essential to maintain user interactions in a state of peak throughput - often referred to as a ’flow state’ (Beck and Csikszentmihalyi, 1990) - for as long as possible. To this end we designed the interface to minimise the occurrence of any full-page reload events, and to minimise any lag while loading the interface.


SENTimental - Cymraeg / Saesneg

ID Astudio: 0001-caab

Ystyriwch y testun isod:

Bore da, Annwen Morgan, prif weithredwraig Cyngor Sir Ynys Môn.

Defnyddiwch y llithrydd isod i aseinio pa deimlad cyffredinol sydd gan y testun hwn a gwasgwch y botwm Anfon, neu dewiswch Skip i symud i destun newydd.

Niwtral

Crëwyd gan John Vidler ar gyfer prosiect SmolLM fel rhan o grŵp ymchwil UCREL yn Ysgol Cyfrifiadura a Chyfathrebu, Prifysgol Caerhirfryn

Defnyddir cwci gwe i storio ID Astudio'r defnyddiwr i'w gadw rhwng sesiynau.

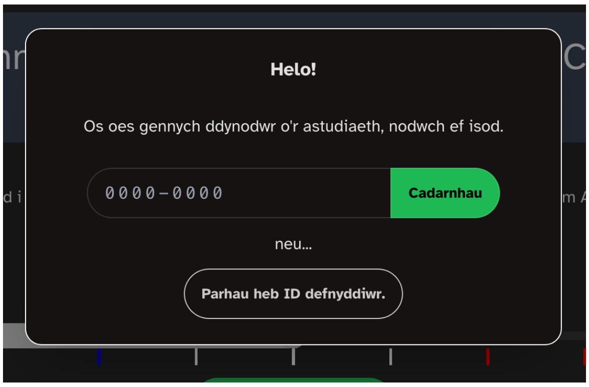
Figure 2: The desktop UI for SENTimental. All text in the interface can be swapped between Welsh and English and the application includes translation files for both which can be customised to match the requirements of the study or language.

While the focus on the design was to facilitate usage in mobile browsers, each design element was also created according to responsive design (Yousaf et al., 2019) patterns, smoothly enabling the transition to larger screen formats without significantly changing the layout to maintain user familiarity at all resolutions and configurations. The SENTimental user interface on a desktop browser (Firefox) can be seen in Figure 2 and includes all the elements seen in the mobile interface but rescaled for the available space.

3.2 Identifying Users

While accepting responses from any interested citizen scientist or volunteer is useful, it is valuable to be able to identify individual users as having been specifically asked to take part in a study; to this end we include the option to sign into SENTimental with a ’Study ID’ (See Figure 3). This is an 8-character code including a checksum to validate entry which if correctly entered then segregated the users data from the general storage code (0000-0000 for unauthenticated users) allowing their input to be easily identified.

The first 4-digits generate a second 4-digit code which is used to validate the user against the second 4-digits in the ID, reducing the number of valid identifier combinations but allowing the system to validate the identifier without any additional information. This identifier is not cryptographically secure, but it being so is not required as each identifier’s data is stored independently, and once identified as being from a malicious source the entire identifier can be discarded if required.


Helo!

Os oes gennych ddynodwr o'r astudiaeth, nodwch ef isod.

Figure 3: The study ID entry dialog; rather than using a traditional username and password combination we elected to use an 8-digit code, which both identifies a user, and acts as a checksum against spam.

Once the user has identified themselves with this Study ID on a given browser platform, it is stored in the browser’s local storage, such that subsequent visits automatically retain the identifier, further reducing any barrier to starting the task.

This identification method was chosen over a traditional login or even a ”passwordless” scheme as it proved to have the least friction when first accessing the tool, as a user can be simply supplied with an 8-character code, enter this and begin, requiring no email or other side-channel authentication method. While we do acknowledge that this method may present a way for unauthorised users to gain access to record annotations, we do not believe this to be a large threat as the built-in checksum for the Study ID prevents blindly entering codes, and the web server instance is configured to limit bursts of activity to further mitigate brute-force attempts.

3.3 Minimise Data Transfer

We have designed SENTimental to require only minimal data transfer to operate; while it does use the React[2] framework to present a component-based architecture and that does cause some filesize overheads, once the basic page is loaded into the browser, the software then only has to retrieve new prompts and report back votes. Both of these are minimal data-only network transfers.

As most of the page is therefore static, this allows the browser to heavily cache the application, reducing load times further and facilitates usage in less-than-ideal network and system conditions (low bandwidth, low performance hardware). This, in turn, maximises the number of devices and situations where the application can be used.

3.4 Data Formats and Storage

SENTimental uses JSON-based records specifically in the JSON Lines format (JSONL)[3], which can be easily processed using widely adopted tools, including many of the HuggingFace libraries for handling datasets. When used to record participant responses, JSONL is advantageous because it allows for the storage of structured data in a line-by-line format allowing SENTimental to simply append new results directly to a storage file without reading the existing data into memory or using a database storage engine.

Similarly, while formats like XML, YAML, and TEI (Text Encoding Initiative) offer their own advantages, they are not as suitable for our needs, notably in terms of speed. Our application is developed in JavaScript which natively supports reading and writing JSON data and it is this native compatibility allows for faster processing and manipulation of data compared to other formats. YAML may be considered for future development due to its overlapping feature set with JSON and its capability to handle log-format storage, our current focus remains on leveraging the efficiency and simplicity of JSON-based records.

```
{
  "en": "It’s let people down, ...",
  "cy": "Mae wedi siomi pobl, ...",
  "negative": 0,
  "positive": 0,
  "neutral": 0,
  "score": 0,
  "label": "neutral",
  "_index": 1202,
  "value": 0.1,
  "vote_language": "cy"
}
```

Figure 4: One row from the annotation storage file. Note that the en and cy fields have been truncated to fit in this paper format and the JSON has been expanded to be more human readable. In the storage files this is a condensed to a single line such that one row is one line for ease of subsequent processing.

Each record in our JSONL format follows the structure described in Figure 4; although in the example presented here the data values have been truncated to fit this format. Each language is keyed by its ISO country code (in this case, ’en’ for English and ’cy’ for Welsh), along with which language the user was viewing the page in when they submitted their vote in the ’vote_language’ field.

The ’value’ field contains the fractional sentiment value from the user, which compares to the ’score’ field as generated by either the English language sentiment analysis, or the new model’s own ’score’ when running tests with the associated scripts. As many of the sentiment analysis tools also offer distinct positive, neutral and negative certainties we additionally include them here for later analysis, although they are not required for our training and testing stages, but act just as an opportunity for further checks.

The only remaining field is ’_index’ which refers to the position in the prompts dataset that SENTimental has been using to retrieve prompts to present to the user, and is used internally to track any issues with these prompts (encoding issues causing render issues on the webpage, for example).

4 Evaluation

To evaluate SENTimental, we generated a training set for a new LLM from a bilingual corpus that incorporates sentiment annotations from existing

[2] https://react.dev/ - React; ”The library for web and native user interfaces”, retrieved May 2025

[3] https://jsonlines.org/ - ”Documentation for the JSON Lines text file format”, retrieved May 2025

English tools and use these to inform the Welsh equivalent texts. For SmolLM we used the Welsh Senedd Open Data covering bilingual transcripts from nearly all sessions in the Senedd as these have already been translated and verified by human translators so we have a high confidence in the accuracy of the translations.

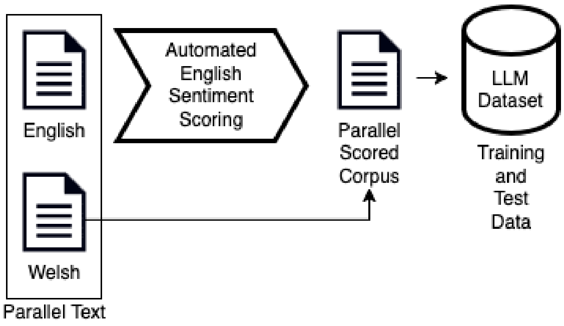


Figure 5: The process of taking the bilingual raw sources and machine annotating only one language so the combined data could be used to train our LLM.

Using the Asent (Enevoldsen, 2022) English sentiment analyser we generated annotations for each of the paired texts then combined these into a single dataset which we then used to train an LLM (See Figure 5). For this process, we selected the ”distilbert-base-uncased” model as our baseline and fine tuned it with our annotated values and Welsh text using the UCREL Hex compute service (Vidler and Rayson).

This same dataset was also used to make an initial test dataset by randomly sampling from the same raw data to create an independent but structurally identical dataset for evaluation.

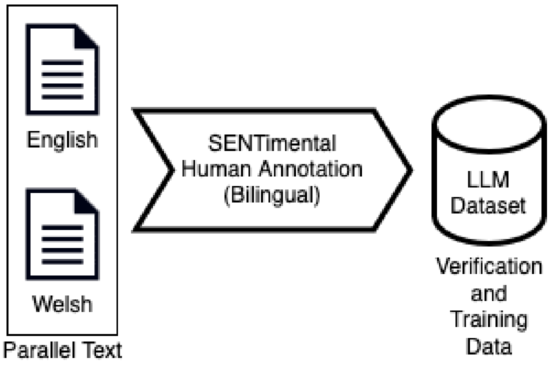

English
Welsh
Parallel Text
SENTimental Human Annotation (Bilingual)
LLM Dataset
Verification and Training Data

Figure 6: Generation of the human-annotated test dataset from the SENTimental response data. As the formats for our original training data and SENTimental are nearly identical it is trivial to use the SENTimental responses directly as a test dataset for the LLM.

Once we trained the model we could then take the responses gathered by SENTimental and compare these human annotations to the machine generated equivalents to assess how effective our model was (See Figure 6).

4.1 Results

While this was only a short study and only had a small number of respondents we still managed to achieve over 4500 individual responses for a total of 109,019 words of Welsh text reviewed. One reviewer was particularly engaged and achieved over 2000 responses alone.

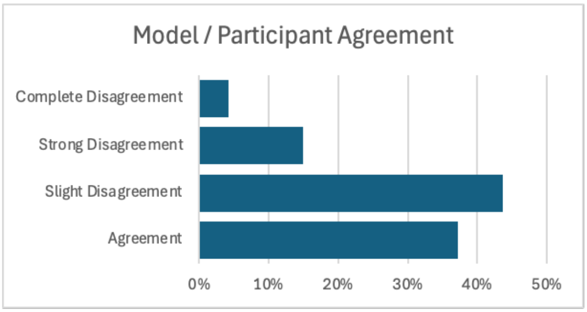


Figure 7: The level of agreement between the LLM and our human reviewers. Agreement is any annotation that perfectly aligned, Slight Disagreement is a 1 category misalignment, Strong Disagreement a 2 category misalignment and Complete Disagreement is any further misalignment. Taking the top two categories here (up to 1 category of misalignment) we see an 80% correlation between the model and human assessment.

Performing the verification process, as described in Section 4, we can compare the results from human assessments and machine assessment, and doing so we see an 80% agreement within a 1-point misalignment (whereby the model ranks a test as ’neutral’ but human annotators rank it as ’slightly negative’ for example).

This demonstrates that we can both accurately predict the sentiment of Welsh text with the LLM, and also that we can describe this success in terms relative to human annotation through our tool SENTimental. To further fine-tune this model, we may also now use the results from the human annotation directly with the existing training scripts to steer the LLM rankings closer to those as defined by our annotators.

5 Deployment with Docker

SENTimental has been written to support simplified deployment through the use of Docker and

Docker Compose. This deployment makes it easy to set up and run the application in a variety of environments, and a minimal setup requires only a list of prompts, and any responses collected will be logged under an anonymous user ID, providing a basic level of functionality for data collection.

For more complex studies or deployments, SENTimental also offers the ability to generate specific user IDs or study IDs, allowing for more fine-grained tracking and analysis of participant responses. The web service can generate small batches of valid IDs, which can then be distributed to participants, enabling them to identify themselves or the study they are part of when accessing the web interface. This feature provides a high degree of flexibility and customization, making it easier to integrate SENTimental into a wide range of research designs and studies.

The build files and source code for SENTimental are freely available on Github at `https://github.com/UCREL/SmolLM`, allowing developers and researchers to access, modify, and extend the application as needed. With Docker installed and configured on a host machine, the application can be deployed in just a few minutes, making it easy to get started with data collection and analysis and can be deployed on a workstation by similar means using Docker desktop or a similar container runtime application. By providing open access to the source code and build files, the developers of SENTimental aim to facilitate collaboration, customization, and community-driven development, ultimately contributing to the advancement of sentiment analysis and natural language processing research.

6 Further Development

In section 3.1 we describe our efforts in SENTimental to maintain a flow state for the user while they use the tool, but there are additional ways that this could be enhanced should the tool continue to be developed. One such way this can be achieved is by making the task more enjoyable and immersive such as through 'gamification' (Dehganzadeh and Dehganzadeh, 2020), which involves incorporating game design elements and mechanics into the task. Language-learning tools like Duolingo[4] and others have successfully implemented gamification, using features such as 'scores', 'levels', and rewards to motivate users and track their progress.

Similarly SENTimental could be enhanced by introducing gamification elements, such as a 'streaks' metric, which rewards users for consistently inputting ratings over a prolonged period. The longer the user continues to input ratings, the higher their continuous 'streak' score becomes. However, this approach may have unintended consequences, such as creating incentives for users to rush through prompts to maintain their streak, potentially compromising the accuracy of the results.

To mitigate this risk, it's crucial to strike a balance between encouraging user engagement and ensuring the quality of the input. One possible solution is to implement enforced breaks, similar to those used in language-learning applications, which allow users to pause and resume their progress without losing their streak. This can help prevent over-optimization of the pseudo-gameplay and maintain a healthy balance between user engagement and data quality.

Additionally, we briefly explored a flipped capture method for SENTimental, whereby the existing predictions from the language model were shown to a user and they were simply asked if they agree or disagree with the assessment. While this was ultimately dropped from the design it may be a useful method for other studies.

7 Open Model and Open Data

The datasets and model here are released on the UCREL HuggingFace pages at `https://huggingface.co/ucrelnlp`, and can be freely used as part of other tools and analysis. We welcome comments and improvements from the community.

Acknowledgments

SENTimental was produced as part of the "A Small Welsh Language Model Pilot for Sentiment Analysis Testing 2024-25" sponsored by the Welsh Government, 2025. This project was part of a collaboration between Cardiff and Lancaster Universities as part a range of collaborative UKRI and Welsh Government funded projects. More details can be found via the Welsh Digital Grid website `https://digigrid.cymru/`. We also acknowledge the work of the manual annotators labelling the corpus data.

[4] `https://www.duolingo.com/` - The Duolingo Homepage, retrieved May 2025

References

Lawrence A. Beck and Mihal Csikszentmihalyi. 1990. Flow: The psychology of optimal experience. *Journal of Leisure Research*, 24(1):93–94.

Hossein Dehganzadeh and Hojjat Dehganzadeh. 2020. Investigating effects of digital gamification-based language learning: a systematic review. *Journal of English Language Teaching and Learning*, 12(25):53–93.

Kenneth Enevoldsen. 2022. Asent: Fast, flexible and transparent sentiment analysis.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Ying Lin, Cash Costello, Boliang Zhang, Di Lu, Heng Ji, James Mayfield, and Paul McNamee. 2018. Platforms for non-speakers annotating names in any language. In *Proceedings of ACL 2018, System Demonstrations*, pages 1–6.

Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Cheri Mullins. 2015. Responsive, mobile app, mobile first: untangling the ux design web in practical experience. In *Proceedings of the 33rd Annual International Conference on the Design of Communication*, SIGDOC '15, New York, NY, USA. Association for Computing Machinery.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

John Vidler and Paul Rayson. UCREL - Hex; a shared, hybrid multiprocessor system. `https://github.com/UCREL/hex`. Accessed: 2025.

Nazish Yousaf, Wasi Haider Butt, Farooque Azam, and Muhammad Waseem Anwar. 2019. A systematic review of adaptive and responsive design approaches for world wide web. In *Advances in Information and Communication Networks*, pages 704–717, Cham. Springer International Publishing.