Hostomytho: A GWAP for Synthetic Clinical Texts Evaluation and Annotation

Nicolas Hiebel[1], **Bertrand Remy**[2], **Bruno Guillaume**[2], **Olivier Ferret**[3], **Aurélie Névéol**[1], **Karën Fort**[4]
[1]Université Paris Saclay, CNRS, LISN, Orsay, France
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[3]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[4]Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, France
[1]{nicolas.hiebel, aurelie.neveol}@lisn.upsaclay.fr
[2]{bertrand.remy, bruno.guillaume}@inria.fr
[3]olivier.ferret@cea.fr, [4]karen.fort@loria.fr

Abstract

This paper presents the creation of Hostomytho, a game with a purpose intended for evaluating the quality of synthetic biomedical texts through multiple mini-games. Hostomytho was developed entirely using open source technologies both for internet browser and mobile platforms (IOS & Android). The code and the annotations created for synthetic clinical cases in French will be made freely available.

Keywords: GWAP, Text Generation, Evaluation, Clinical Texts, Synthetic Texts, French

1. Introduction

One of the most common hurdles in Natural Language Processing (NLP) is the lack of specific resources, whether it be task-specific resources, domain-specific resources, or both. A major challenge for clinical NLP is the lack of shared clinical corpora in languages other than English (Névéol et al., 2018). One potential approach to address this problem is to generate new corpora automatically. The generated corpus should share as many characteristics as possible with the natural corpus, without simply copying it. Thus, evaluating the quality of the generated corpus is crucial. In this work, we generate synthetic clinical texts from real clinical corpora.

We decided to develop a Game With A Purpose (GWAP) to help with the evaluation of the synthetic texts as GWAPs have been proven to be a promising alternative to traditional human annotation.

In this paper we present Hostomytho, a game made for manually evaluating synthetic clinical documents. The game is multi-platform and developed using open source technologies.

The main contribution of this work is an open source game platform set-up to collect linguistic resources to address the following research questions:

- Can a GWAP be a suitable interface for the evaluation of text generation?
- Is medical training needed for evaluating clinical text?
- Can high quality annotations be collected for this complex evaluation task?

2. GWAPs for Language Resources

GWAPs have been used with success for nearly two decades in NLP (Lafourcade, 2007; Chamberlain et al., 2008) to create a wide variety of language resources, from part-of-speech tags (Madge et al., 2019) to word-sense labels (Venhuizen et al., 2013). They proved efficient, even on complex tasks that require training, like dependency syntax annotations (Guillaume et al., 2016). Moreover, they do not present the same ethical issues as microworking crowdsourcing (Fort et al., 2011). To our knowledge, there has been yet no GWAP developed to validate and annotate specialized synthetic texts.

3. Evaluation of Natural Language Generation

Natural Language Generation (NLG) is an area of NLP that has grown in popularity with the advent of pre-trained large language models. A major challenge when doing NLG is the evaluation part. Existing automatic evaluation methods are limited (Novikova et al., 2017) and new measures are often put forth to address those limitations (Frisoni et al., 2022; Pillutla et al., 2021).

It is still accepted that manual evaluation is the best way to evaluate the quality of automatically generated text despite the creation of new automatic metrics. However, manual evaluation of text generation also comes with a number of challenges (Gehrmann et al., 2023; Celikyilmaz et al., 2021). Assessing the overall quality of long sequences of text makes it difficult to maintain consistence during evaluation. This is further exacerbated by the

Games and NLP @LREC-COLING-2024, pages 14–20

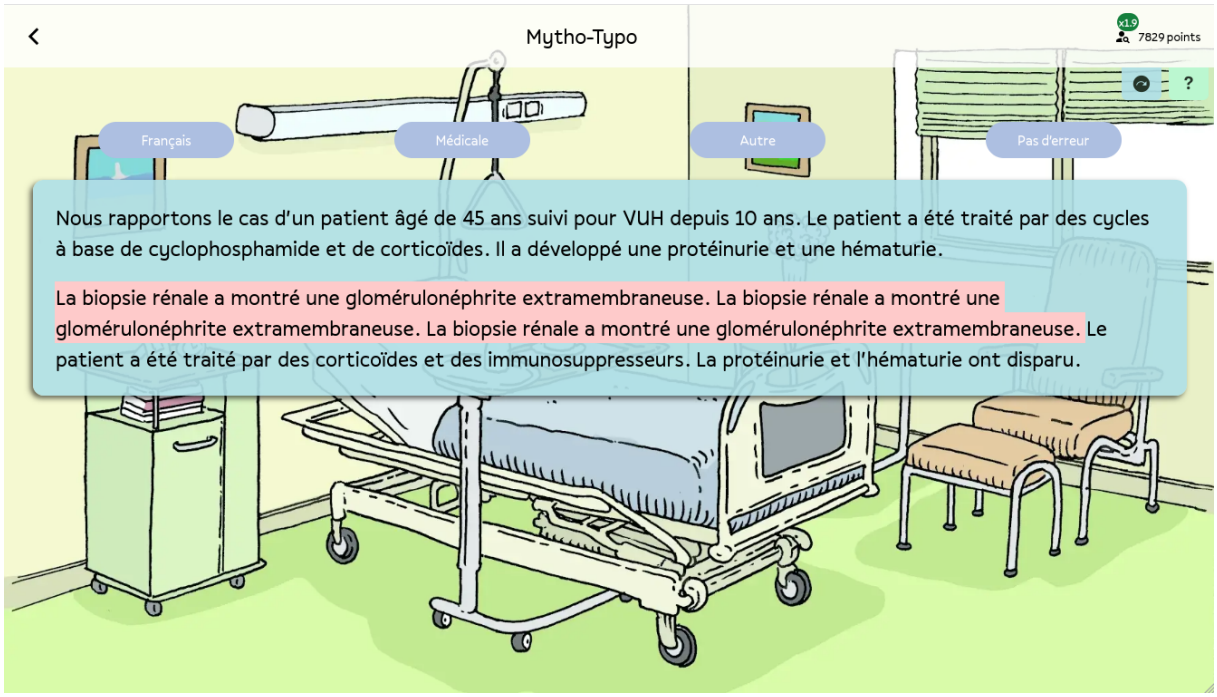

Figure 1: Mytho-Typo: an error type specification game.

broad definition of the terms used to gauge text quality (e.g. fluency, coherence) (Howcroft et al., 2020). Additionally, reviewing long passages of text can quickly become tedious, even more so when it comes to clinical documents which may contain bad outcomes.

4. Evaluating Synthetic Clinical Documents

4.1. Corpus

The generated texts we use in `Hostomytho` come from models trained on French clinical cases (Hiebel et al., 2023). The training documents were collected from the CAS corpus (Grabar et al., 2018) and the French part of the E3C corpus (Magnini et al., 2020), two freely available corpora. An example of a generated clinical case can be seen in Example (1).

(1) *Il s'agissait d'un patient de 50 ans, sans antécédents pathologiques particuliers, admis aux urgences pour des douleurs épigastriques aiguës associées à une distension abdominopelvienne évoluant depuis deux jours. L'examen clinique trouvait un patient en assez bon état général (Apgar: 10/10). Le bilan préopératoire objectivait une fonction rénale normale et la CRP était à 12 mg/l.*[1]

We also added some real clinical cases and some irrelevant documents in the game in order to control the quality of the annotations. The irrelevant document are taken from the corpus Est Républicain (ATILF and CLLE, 2020), a journalistic corpus. We selected documents with vaguely medical content. We expect that there will be no annotation error on the real clinical cases and that irrelevant texts will be detected as such.

4.2. Grammar, fluency and clinical coherence

Several types of errors can be found in automatically generated clinical texts. Some of them might be easy to spot, for example when the text has clear grammatical or fluency issues.

However, most of the time, recent language models manage to generate fluent text. Working with data from a specialized field such as the clinical domain comes with additional challenges. Medical knowledge might be required to spot clinical incon-

[1] Translation into English: *The patient is a 50 years old male admitted to the emergency room with a 2-day history of acute epigastric pain associated with abdominopelvic distension. His past medical history was unremarkable and he was generally in a good state of health (Apgar: 10/10). The preoperative workup showed normal renal function and CRP was at 12 mg/l.*

sistencies. Those are often due to the combination of several elements in the text, unproblematic when taken separately. Looking at example (1), a 50 year old patient is associated with an Apgar score, a method intended to evaluate the health of newborn babies.

Many error typologies exist (Howcroft et al., 2020). Looking at the generated texts and for simplicity, we identify three main types of error:

- grammatical errors: it can be non-existing words or ungrammatical constructions;
- fluency errors: the text seems to be a sequence of unconnected parts or has repeated parts;
- clinical inconsistency: the text is grammatical and fluent, but contains clinically contradictory evidence.

4.3. Divide and Conquer Approach

As mentioned in Section 4.2, evaluating the quality of generated text is a complex multidimensional task. Trying to evaluate a text in detail in one go is intellectually demanding and can quickly become tedious and prone to errors.

Bernstein et al. (2015) proposed the "find-fix-verify" workflow for a writing assistance service to reduce cost and to ensure annotation quality. The task is decomposed in three stages involving different annotators: (i) annotators identify an area of the text that could be improved, (ii) annotators propose modifications to improve a previously identified area and (iii) annotators validate or invalidate the candidate modifications.

We also decided to decompose the evaluation process in different tasks, both to ease the mental burden of players and to have more control over the different types of annotations.

`Hostomytho` currently includes two games. The first game consists in assessing the plausibility of a given text on a scale of five labels ranging from highly implausible to very plausible. The player can select a span of text if an error is present. The second game exploits the results of the first game. The player must classify the type of errors that were annotated. An example can be seen on Figure 1. The player has to choose between four options given a text where the annotated error is highlighted. We've kept the number of options low for simplicity. They are as follows:

- ***Français***: French, for grammatical and fluency errors;
- ***Médicale***: medical, for medical inconsistencies;
- ***Pas d'erreur***: no errors, when the span of text was mistakenly annotated as error.
- ***Autre***: others, for errors that do not fall into other categories.

On the example text of Figure 1, the error highlighted is the repetition of the same sentence three times in the text. This is a fluency error that should be classified with the label *Français* (French).

Each game starts with a tutorial that helps the player understand the current task. At the end of the tutorial, the player gets to practise on a sample of texts for which gold standard annotations are available. This helps us ensure that the player understood the task well enough before starting the annotation of new texts.

4.4. Control over Annotation Quality

We plan to control the quality of the annotations in two different ways. First, we will check agreement between players on each task by sharing some samples between players. This will also help us assess the difficulty of the task.

Second, we assign a neutral reliability score to every player on account creation (50 on a scale from 0 to 100). We will occasionally give players control samples where the correct answer is known and the reliability score will increase or decrease depending on the players' answers on the control samples.

Annotations given by a player with a high reliability score will carry more weight that those given by a player with a low reliability score.

5. Player types

Players can find satisfaction in different elements depending on their profile and several player taxonomies have been proposed (Bartle, 1996; Tondello et al., 2016). Here's a brief description of the four types of players according to Bartle (1996):

- *Achievers* enjoy accomplishing different things in the game;
- *Explorers* enjoy discovering every parts of the game;
- *Socializers* enjoy interacting with other players;
- *Killers* enjoy attacking other players.

We will link the types of players with the game elements in Section 6.2.

6. Presenting `Hostomytho`

6.1. The Universe

The task of differentiating between real clinical documents and synthetic clinical documents is the main motive for `Hostomytho`'s universe.

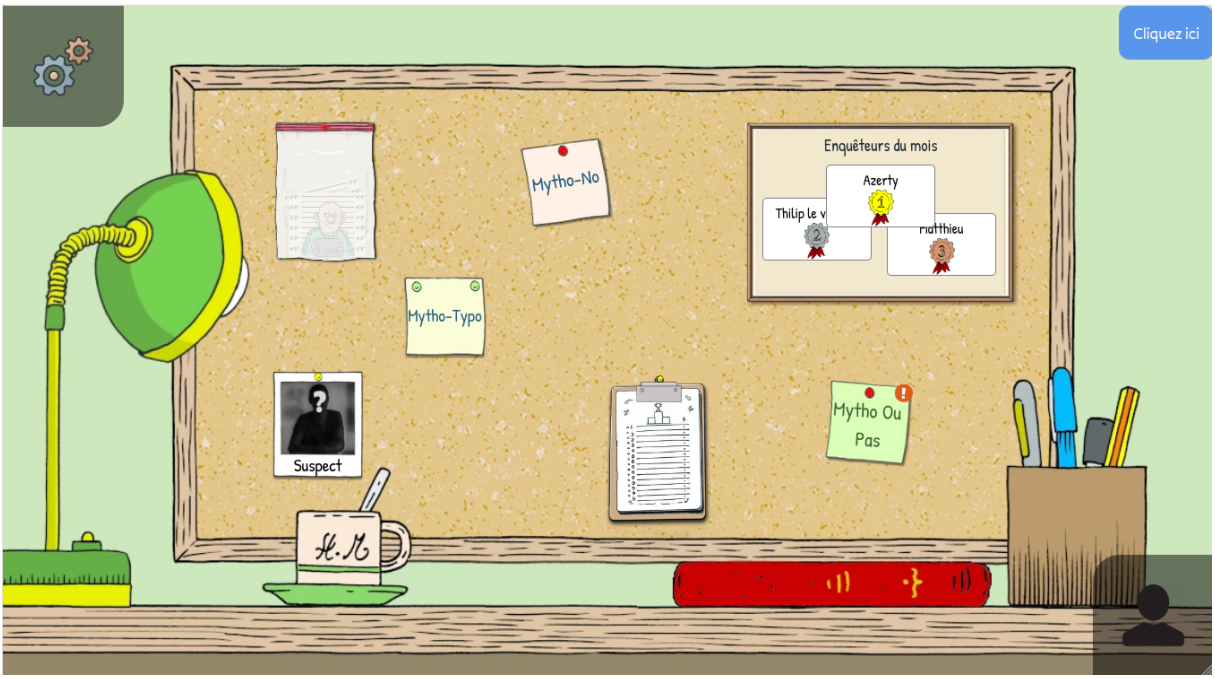

Figure 2: Main menu with leaderboard on the top right corner of the cork board.

We are casting this as a mystery scenario. In the game story, criminals have escaped and found refuge in a hospital. They hide by pretending to be doctors. The player takes the role of an investigator mandated to unmask the criminals hidden in the hospital. The player investigates by analyzing the clinical documents produced by the hospital and looking for potential errors. As the player gathers more clues, the investigation progresses, leading to the arrest of increasingly tough suspects.

6.2. Game Elements

Game elements in `Hostomytho` focus on earning points and progressing the investigation of criminals. We offer several game mechanics to meet the needs of different types of players, as identified in (Bartle, 1996).

6.2.1. Leaderboard and Ranking

The point system represents the player's overall progression. Points are acquired by playing the different mini-games. Players can keep track of their rank by checking the leaderboard. This encourages the players to play more to move up in the rankings.

In addition to a global leaderboard, we added a special spot on the main menu for the best investigators of each month. With a monthly ranking, players should come back regularly to be on top of the ladder.

Figure 2 shows the main menu of `Hostomytho`. The monthly top three investigators are displayed in the top right corner of the cork board. Players having a chance to have their username and avatar displayed in the main menu for everyone to see should be motivating to play for more points.

This part of the game focusing on being the best player should appeal to achievers.

6.2.2. Investigation and Achievements

The player's main goal in `Hostomytho` is to arrest as many criminals as possible. Players can try to catch the criminal they're currently tracking at any time during the game. Each arrest has a certain chance to succeed depending on the player's "certainty". The certainty score can be increased by completing more tasks. We hope this system will encourage players to complete more tasks in order to maximize their chances of success.

The player tracks one criminal at a time. With each arrest, they move on to the next criminal and will progressively encounter criminals harder to catch. We want to give players the feeling that they're making progress in the investigation so that they'll want to go further.

We also use achievements as a way of motivating players and rewarding them for playing the game regularly. Achievements can be obtained by completing various objectives like arresting criminals or playing the game for several days in a row. Besides, unlocking achievements increases the rate

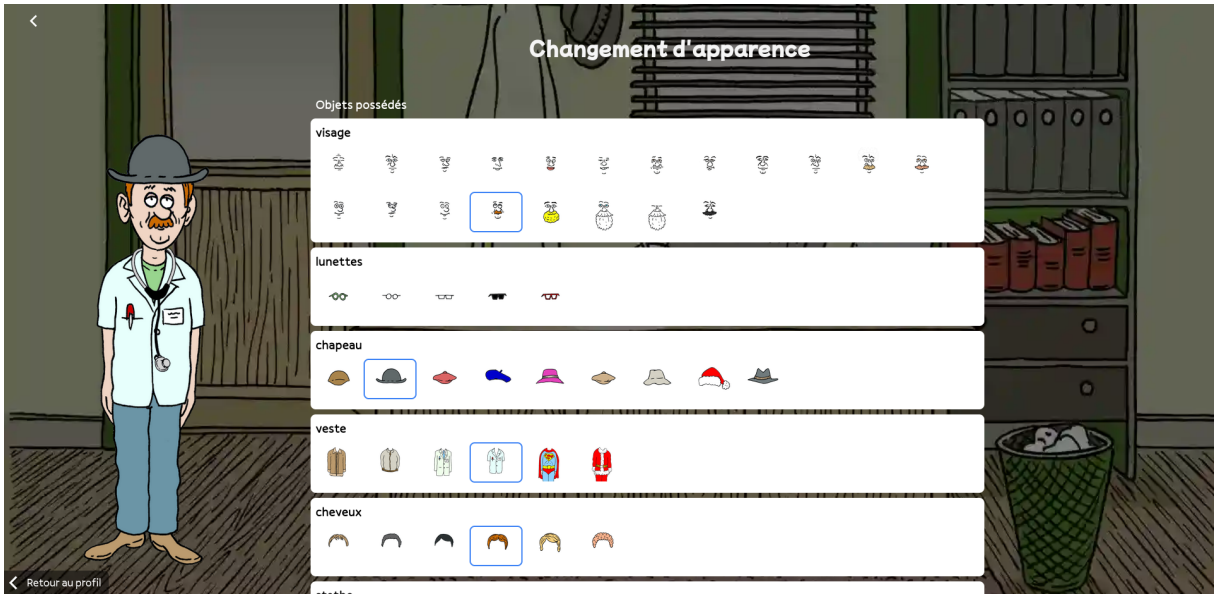

Figure 3: Character customization menu.

at which players earn points, making achievements not only satisfying but also useful for progression.

Progressing in the investigation and earning achievements should also appeal to achievers. In addition, discovering new criminals should appeal to explorers.

6.2.3. Skins and Customization

In `Hostomytho`, each player uses an avatar that represents their investigator. The avatar can be customized with hair, hats, clothes, and other accessories. Players will regularly unlock new items for customization when earning points. Figure 3 shows an example of the character customization menu (with some items already unlocked). Players may find satisfaction in personalizing their own investigator's avatar with the different items they unlocked. Some items are less common than others and discovering new ways of customizing the avatar might excite the curiosity of players.

Collecting customization elements should appeal to explorers.

6.2.4. Covering All Player Profiles

At the time of writing, the game elements of `Hostomytho` are primarily aimed at achievers and explorers. We plan to add game elements that will meet the needs of the other types of players.

For the socializers, we plan to add a friend system so that players can compare their scores with those of their friends. In addition, players will be able to group in companies of investigators and work together to place their company at the top of the company leaderboard. Finally for the killers, who like to attack other players, we plan to add the possibility of playing the role of a criminal. In this role, the player will be able to select a generated (fake) text from several generated texts. The selected text is then presented to an investigator. If the investigator finds no error in the text, the criminal will have succeeded in deceiving the investigator and will earn points.

7. Conclusion and Future Work

`Hostomytho` development is already well underway. At the time of writing, two mini-games are already available and the game is being tested for bugs and feedback. We are planning to add more games to obtain different annotations, which should be facilitated by the reliable base we already have. New games should include negation detection, hypothesis detection and condition detection. These annotations could help improve existing information extraction tools in the clinical domain by providing a more detailed representation of the clinical case. The code for `Hostomytho` is completely open-source and will be made available when the game is stable.

We plan to annotate two sets of generated texts. The first set will be generated with models trained on the freely available corpora or clinical cases in French CAS and E3C. For the second set, we plan to train the models on non-shareable medical reports in French. We will wait for the committee's approval to add the texts to the game. The annotations of the texts generated from the freely available corpora will also be freely available. We will also wait for the committee's approval to share the anno-

tations on the second set of texts, generated from the private data.

8. Acknowledgements

This work has received funding from the French "Agence Nationale pour la Recherche" under grant agreement CODEINE ANR-20-CE23-0026-01.

9. Bibliographical References

R. Bartle. 1996. Hearts, clubs, diamonds, spades: Players who suit MUDs. *The Journal of Virtual Environments*, 1(1).

Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM*, 58(8):85–94.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey. *arXiv:2006.14799 [cs]*.

Jonathan Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase Detectives: a web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz, Austria.

Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2):413–420.

Giacomo Frisoni, Antonella Carbonaro, Gianluca Moro, Andrea Zammarchi, and Marco Avagnano. 2022. NLG-metricverse: An end-to-end library for evaluating natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3465–3479, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Bruno Guillaume, Karën Fort, and Nicolas Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition. In *Proc. of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand.

Chris Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2019. Making text annotation fun with a clicker game. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG'19, pages 77:1–77:6, New York, NY, USA. ACM.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828.

Gustavo F. Tondello, Rina R. Wehbe, Lisa Diamond, Marc Busch, Andrzej Marczewski, and Lennart E. Nacke. 2016. The gamification user types hexad scale. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY'16, pages 229–243, New York, NY, USA. Association for Computing Machinery.

Noortje Joost Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 397–403, Potsdam, Germany. Association for Computational Linguistics.

10. Language Resource References

ATILF and CLLE. 2020. *Corpus journalistique issu de l'Est Républicain*. PID http://redac.univ-tlse2.fr/corpus/estRepublicain.html. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.

Grabar, Natalia and Claveau, Vincent and Dalloux, Clément. 2018. *CAS: French Corpus with Clinical Cases*. Association for Computational Linguistics. PID https://deft.lisn.upsaclay.fr/2020.

Bernardo Magnini and Begoña Altuna and Alberto Lavelli and Manuela Speranza and Roberto Zanoli. 2020. *The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases*. CEUR-WS.org. PID https://github.com/hltfbk/E3C-Corpus.