Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 856–866
Marseille, 20-25 June 2022

`Textinator`: an Internationalized Tool for Annotation and Human Evaluation in Natural Language Processing and Generation

Dmytro Kalpakchi, Johan Boye
Division of Speech, Music and Hearing
KTH Royal Institute of Technology
Stockholm, Sweden
{dmytroka,jboye}@kth.se

Abstract

We release an internationalized annotation and human evaluation bundle, called *Textinator*, along with documentation and video tutorials. Textinator allows annotating data for a wide variety of NLP tasks, and its user interface is offered in multiple languages, lowering the entry threshold for domain experts. The latter is, in fact, quite a rare feature among the annotation tools, that allows controlling for possible unintended biases introduced due to hiring only English-speaking annotators. We illustrate the rarity of this feature by presenting a thorough systematic comparison of Textinator to previously published annotation tools along 9 different axes (with internationalization being one of them). To encourage researchers to design their human evaluation before starting to annotate data, Textinator offers an easy-to-use tool for human evaluations allowing importing surveys with potentially hundreds of evaluation items in one click. We finish by presenting several use cases of annotation and evaluation projects conducted using pre-release versions of Textinator. The presented use cases do not represent Textinator's full annotation or evaluation capabilities, and interested readers are referred to the online documentation for more information.

Keywords: annotation tool, human evaluation tool, natural language processing, natural language generation

1. Introduction

Large-scale pretrained language models, also called foundation models, have brought substantial advances to the Natural Language Processing (NLP) field. Indeed, currently the dominant approach for many NLP tasks is to fine-tune a foundation model on relatively small amounts of task-specific annotated data (Bommasani et al., 2021, Section 2.1.2). These recent developments introduce two major shifts to the field. The first shift is from focusing on large variable-quality crowdsourced annotated datasets back to relatively small high-quality datasets. The second shift is from focusing mostly on methods to focusing mostly on evaluation. In this paper we attempt at addressing both shifts by presenting a tool, called `Textinator`[1], suitable for annotating data for a wide variety of NLP tasks, as well as facilitating human evaluation.

Addressing the first shift in more detail, we believe that first and foremost, high-quality data should have correct and *consistent* annotations. Such annotations are easier to obtain with the help of domain experts, who are, evidently, more expensive to hire than crowdworkers. This is why annotation tools should have a low entry threshold and require no specialized training. Secondly, high-quality data should limit biases only to those necessary for exploring the concept at hand. For instance, factual question answering datasets will necessarily be biased towards wh-questions, but will usually not contain why-questions.

One very subtle (and usually unintended) way of introducing bias in the data is recruiting annotators based on their knowledge of English, since an annotation tool at hand offers UI only in English. Unless data collection happens in an English-speaking country, knowledge of English correlates positively with high socio-economic status (McCormick, 2013). Such bias might not always be desirable, especially for the studies aimed at language perception. For instance, researchers interested in studying language toxicity might want to recruit participants with different socio-economic backgrounds to accommodate differences in the perception of toxicity.

The observation above prompts modern annotation tools to be *internationalized and localized* into as many languages as possible. *Localization* refers to the process of providing translations of UI elements into different languages (usually done by translators). The process of preparing software for localization is referred to as *internationalization* (usually done by engineers). If annotations are done for language other than English, localized software will not only allow controlling for the aforementioned type of bias, but will most certainly lower the entry threshold for domain experts. While trained linguists might know the English names for linguistic phenomena (although it tends to vary across the countries), annotating using linguistic terms in their mother tongue will most certainly lower their cognitive load (and thus annotation time).

Addressing the second shift (from focusing mostly on methods to focusing mostly on evaluation), we note that human-in-the-loop evaluation will probably turn out to be critical for natural language generation (NLG)

[1] Both the code and the documentation are available at `https://github.com/dkalpakchi/Textinator` (DOI: 10.5281/zenodo.6497334)

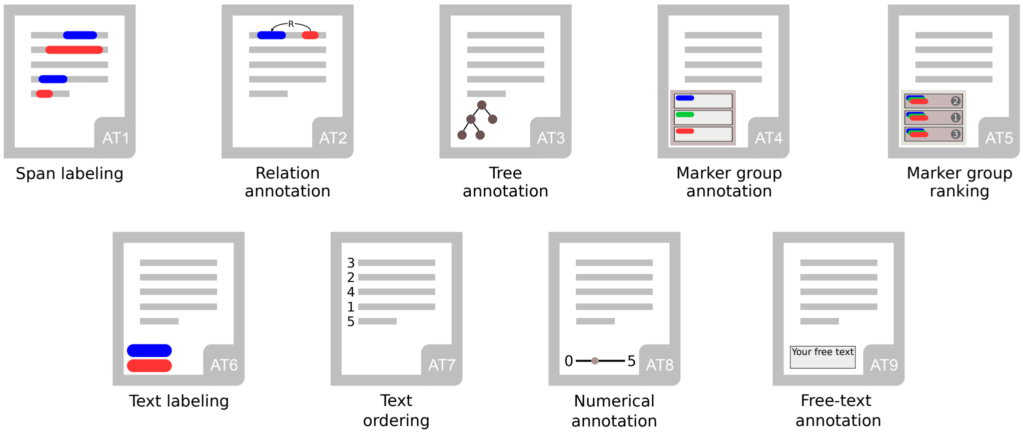


Figure 1: Task types for a within-document annotation

tasks (Bommasani et al., 2021, Section 4.4). Indeed, while fine-tuned foundation models are capable of generating texts that *appear* coherent and well-structured, it is quite often not the case when looking closer. Dou et al. (2021) encountered a wide range of errors including those related to fact truthfulness, redundancy, and common sense, to name a few. Kirk et al. (2021) report that texts generated with GPT-2 contain occupational biases related to gender and ethnicity. Capturing such errors requires a more in-depth human evaluation of NLG models. We believe that the best way forward is to design human evaluations for NLG models before the actual data collection has even started! Our way of encouraging this is by providing an easy-to-use evaluation tool bundled together with an annotation tool. Indeed, `Textinator` is released in hopes that it will facilitate constructing datasets in many languages and conducting thorough human evaluations faster, at least for some NLG tasks.

2. Comparison to Other Tools

2.1. Annotation Tools

The number of annotation tools has skyrocketed over the years providing a rich variety to choose from. To make the comparison systematic, we have defined a number of axes to compare along. We want to emphasize that we performed the comparison based on the information provided in the published papers (if available) and/or the official documentation (if available). We have neither run any of the tools ourselves, nor looked at the source code, nor contacted the authors. In case the tool has an enterprise version, we evaluate only parts freely available to the public.

Axis 1: Task Type Coverage

We consider only within-document annotation task types, thus excluding between-document annotation tasks (e.g., multi-document summarization) or tasks binding current document with external sources (e.g., entity linking). The exact list of considered task types is provided in Figure 1 and exemplified below.

AT1 Named entity recognition or extractive summarization are instances of *span labeling*: the act of marking a span of words in the text and annotating it with a label representing a concept of interest.

AT2 Pronoun resolution is an instance of *relation annotation*: the act of binding two previously labeled text spans and annotating this binding with a named relation.

AT3 Dependency tree annotation is, evidently, an instance of *tree annotation*. Although trees can be represented as sets of relations, they might require a different UI to speed up the annotation.

AT4 Multiple choice question generation is an instance of *marker group annotation*: the act of marking a number of spans by different labels and submitting them as a unit (or providing free text, instead of labeling).

AT5 Multiple choice question complexity ranking is an instance of *marker group ranking*: the act of ordering previously annotated marker groups with respect to a pre-defined criterion.

AT6 Text classification is an instance of *text labeling*: the act of annotating the whole text with a label, representing the concept of interest.

AT7 Ordering restoration of scrambled sentences is an instance of *text ordering*: the act of annotating each sentence with an order of appearance in the original text.

AT8 Sentiment annotation is an instance of *numerical annotation*: the act of associating a number (integer or floating-point) with a text/sentence.

AT9 Abstractive text summarization or machine translation are instances of *free-text annotation*: the act of associating a free text with the whole text/sentence.

We have categorized annotation tools based on how many *task types* they support. We consider the task type to be supported if it is possible to annotate *at least one* task of this type. Using the definition above, we have adopted the following categorization.

[✔] "limited", ≤ 3 supported task types;

(✔) "moderate", $3 <$ supported task types ≤ 6;

✔ "extensive", > 6 supported task types.

The comparison results for this axis are presented in the column A1 of Table 1 with extensive details in Table 2 of Appendix.

Axis 2: Internationalization

First and foremost, we assume that *all* tools support Unicode input texts, since all major programming languages provide Unicode support by default. One potential problem might be in displaying right-to-left languages, which is, however, impossible to assess by only reading papers and documentation, hence the assumption. With this assumption in mind, we have assessed the degree of internationalization according to the number of internationalized UI parts. We consider a UI part to be *internationalized* if:

- it is translated into at least 2 languages;
- a clear mechanism of supplying translations of UI elements is documented;
- it is possible for the annotator to switch the language using UI interactions.

For the purpose of this paper, we consider the following UI parts: static UI elements (e.g., menu items), static markers (supplied with the tool), static relations, dynamic markers (created by users on the fly with user-supplied translations) and dynamic relations. One could argue that internationalization of dynamic markers is not a necessity and one could simply provide translations to a local language directly. However, internationalization is crucial if researchers aim to use *the same annotation scheme across languages*, as for example Universal Dependencies (Nivre et al., 2020) do. The annotators creating UD-based part-of-speech datasets would mark nouns as "noun" for English, "substantiv" for Swedish or "іменник" for Ukrainian, whereas all exported datasets would contain one and only label NOUN.

Bearing in mind the points above, we have then discretized the degree of internationalization as follows.

✘ "none", no internationalized parts;

[✔] "limited", 1 internationalized part;

(✔) "moderate", 2 internationalized parts;

✔ "extensive", 3 or more internationalized parts.

The comparison results for this axis are presented in the column A2 of Table 1 with extensive details in Table 3 of Appendix.

Axis 3: Customization

We have assessed the degree of customization according to how flexible one can be with the definition of the annotation tasks. Specifically, we have looked whether the following objectives can be achieved.

C1 Can custom markers for spans of text be defined?

C2 Can the appearance of the markers be customized?

C3 Can custom hotkeys be defined for markers?

C4 Can custom information be associated with markers (e.g., free-text comments)?

C5 Can custom relations be created for markers?

C6 Can custom constraints on the relations be specified?

C7 Can the appearance of relations be customized?

C8 Can custom hotkeys be defined for relations?

C9 Can custom marker groups be defined to be annotated as a unit?

C10 Can the appearance of marker groups be customized?

C11 Can custom constraints on marker groups be specified?

C12 Can the annotation UI layout be customized?

If an objective can be achieved using *only UI interactions*, a tool scored 1 point for the objective. If an objective is possible to achieve through other means of configuration that *do NOT require programming skills*, e.g., via configuration files, a tool scored 0.5 points for the objective. In all other cases a tool scored 0 points for the objective. Based on the final number of points we have defined 4 roughly equal categories.

✘ "none", 0 points

[✔] "limited", $0 <$ points < 4

(✔) "moderate", $4 \leq$ points < 7

✔ "extensive", ≥ 7 points

The comparison results for this axis are presented in the column A3 of Table 1 with extensive details in Table 4 of Appendix.

Axis 4: Annotation Progress Tracking

Tracking annotation progress is vital. This includes, but is not limited to, getting a birds-eye-view on the aspects on the annotated data (e.g., distribution of markers or distribution of marker lengths), tracking the data source progress (e.g., how much of the data was already annotated, if there were any problems with any of the texts) or being able to identify problematic annotation cases marked by the annotators. It is hard to define the exact list of mandatory features for this axis, which is why we simply use the number of relevant features to divide tools into the following categories.

Annotation tool	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	(A8)	(A9)
WordFreak (Morton and LaCivita, 2003)	[✔]	✘	✘	✘	✘	✘	(✔)	[✔]	🖥
MMAX2 (Müller and Strube, 2006)	[✔]	✘	[✔]	✘	✔	✘	(✔)	✘	🖥
BRAT (Stenetorp et al., 2012)	[✔]	✘	(✔)	(✔)	(✔)	✘	(✔)	(✔)	🗄
WebAnno (Yimam et al., 2013)	[✔]	✘	(✔)	✔	✔	(✔)	✔	(✔)	🗄
GATE Teamware (Bontcheva et al., 2013)	[✔]	✘	[✔]	✔	✔	✔	[✔]	[✔]	🗄
INCEpTION (Klie et al., 2018)	(✔)	✘	✔	✔	✔	✔	✔	✔	🗄
AWOCATo (Daudert, 2020)	[✔]	✘	[✔]	✘	(✔)	✘	[✔]	✘	🗄
Doccano (Nakayama et al., 2018)	[✔]	[✔]	[✔]	✘	✘	(✔)	(✔)	✘	🐳
Label Studio (Tkachenko et al., 2020)	[✔]	✘	(✔)	✘	✘	✘	✔	✔	🐳
`Textinator` (ours)	✔	✔	✔	✔	✘	✔	✔	(✔)	🐳

Table 1: Comparison to other open-source tools (presented in the temporal order). A1 - task type coverage, A2 - internationalization, A3 - customization, A4 - annotation progress tracking, A5 - quality assurance, A6 - administration quality, A7 - documentation quality, A8 - intelligent assistance, A9 - deployment mode. The ones in italic do not have an associated paper, but are instead available on GitHub

✘ "none", no features

(✔) "moderate", up to 2 features

✔ "extensive", 3 features or more

The comparison results for this axis are presented in the column A4 of Table 1 with extensive details in Table 5 of Appendix.

Axis 5: Quality Assurance

Ensuring the quality of the collected data is a very time-consuming task, which is why any features geared towards helping to curate the data are of extreme importance. For instance, if multiple people are annotating the same text, checking between-annotator or within-annotator consistency would help. It is hard to define a list of mandatory features, so we use the same scale as for the Axis 4. The comparison results for this axis are presented in the column A5 of Table 1 with extensive details in Table 6 of Appendix.

Axis 6: Administration Quality

Crowdsourcing is a way to reduce the administration burden for a project manager. It should be used for annotation tasks, when it is possible (or even preferable). Thus features integrating an annotation tool with crowdsoucring platforms are of value. However, we do **NOT** consider such features for this axis and evaluation as a whole for two reasons. First, platforms might have changed their API or stopped operations at all. Secondly, annotation tools developed earlier will be at a disadvantage, since crowdsourcing might have been not widespread back then.

In cases when data collection requires domain experts, crowdsourcing is typically not an option and the burden of project administration (e.g., tracking time and paying salaries) lies on the shoulders of a project manager. With that in mind, we defined "administration quality" as the availability of features related to administering the project, but not the annotation process itself. For instance, if annotators are employed at hourly rate, the tool should be able to provide an estimate of the spent time each month. The availability of different roles per user (e.g., annotators, reviewers, translators) helps ensuring that every user has access only to parts of the data and interface they need. Again, it is hard to define the exact list of mandatory features, so we use exactly the same scale as for the Axis 4. The comparison results for this axis are presented in the column A6 of Table 1 with extensive details in Table 7 of Appendix.

Axis 7: Documentation Quality

Well-documented tools are easier to use than undocumented and the more interactivity the documentation provides, the better. We have considered the following criteria in our assessment:

C1 whether text-form tutorials are provided;

C2 whether video tutorials are provided;

C3 whether sandbox demo is provided;

C4 whether source code documentation is provided for those with programming skills wishing to extend the tool.

Based on the final number of fulfilled criteria we have defined the following 4 categories.

✘ "none", no fulfilled criteria;

[✔] "limited", 1 fulfilled criterion;

(✔) "moderate", 2 fulfilled criteria;

✔ "extensive", 3 or more fulfilled criteria.

The comparison results for this axis are presented in the column A7 of Table 1 with extensive details in Table 8 of Appendix.

Axis 8: Intelligent Assistance
Scaling up data collection in a time-efficient manner would be facilitated if the tool included some kind of intelligent assistance to the annotator. We have considered the following criteria in our assessment:

C1 whether annotation suggestions are provided on the fly out of the box by the tool itself;

C2 whether there is clear method to connect external models for suggestions the fly;

C3 whether there is a possibility of pre-annotating texts;

C4 whether custom suggestion models can be trained on already annotated data and then used for automatic annotation;

C5 whether configurable sanity checks are available (e.g., everything needed is provided);

C6 whether active learning is possible;

C7 whether spelling correction is available;

C8 whether grammar correction is available.

Based on the final number of fulfilled criteria we have defined the following 4 categories.

✘ "none", no fulfilled criteria;

[✔] "limited", 1 fulfilled criterion;

(✔) "moderate", 2 or 3 fulfilled criteria;

✔ "extensive", 4 or more fulfilled criteria.

The comparison results for this axis are presented in the column A8 of Table 1 with extensive details in Table 9 of Appendix.

Axis 9: Deployment mode
We distinguish between a standalone desktop application (🖥), a web application that needs installation on a server (🗄) and a dockerized web application (🐳).

2.2. Evaluation Tools

As mentioned previously, human evaluation in NLG is bound to become more important. To the best of our knowledge, currently researchers are utilizing external web-based survey software, such as Google Forms[2], SurveyMonkey[3], Qualtrics[4], etc. However, we find that these services have two major problems. First, they require manual input, which might end up being quite tedious if the task is, say, to evaluate 50 generated texts by 3 different models according to a number of different criteria. Second, researchers based in the EU need to spend extra effort to comply with GDPR (for instance, by ensuring that their surveys are hosted on the EU servers). These two aspects led us to implement a human evaluation tool, called *Textinator Surveys* (TS). TS can be hosted on the university servers after a simple Docker-based installation, thus solving a GDPR issue. TS also allows importing surveys from a JSON file, in which both the survey's configuration and the items themselves can be specified. TS takes care of item randomization and allows associating meta-information (invisible to a human evaluator) with survey items, which might be very useful for a researcher in the analysis afterwards. See more about the capabilities of TS (and Textinator in general) in the documentation[5] or video tutorials[6].

3. Use Cases

Pre-release versions of Textinator were used both for annotating new datasets and conducting human evaluation in a number of projects, which we briefly present below. Note that in all provided screenshots, the UI localization is turned off for the benefit of the reader of this article, whereas annotated texts are provided in their original language.

3.1. Annotation

Ahlenius (2020) annotated a pronoun resolution dataset for Swedish using one of the earliest versions of Textinator (current UI for this task is shown in Figure 2).
Kalpakchi and Boye (2021a) used Textinator for annotating a dataset of multiple choice questions in Swedish (see Figure 3 for the UI configuration).
Tengvall (2020) annotated a question answering dataset in Swedish using Textinator's configuration similar in Figure 3, but without the red "Distractor" markers.

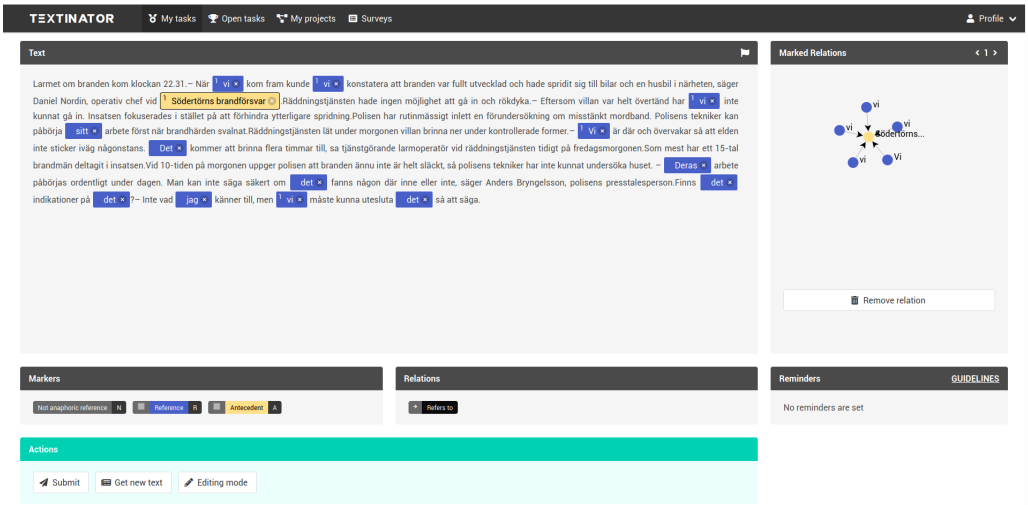
3.2. Human Evaluation

Lindqvist (2021) used Textinator for evaluating the quality of automatically generated question paraphrases in Swedish. Both extensive HTML guidelines (see Figure 5) and the survey items themselves (see Figure 4) were provided using the "Import from JSON" functionality.
Kalpakchi and Boye (2021b) used Textinator for evaluating the quality of generated question-answer pairs in 4 different languages (see Figure 6).
Kalpakchi and Boye (2021a) used Textinator for evaluating the quality of generated distractors in Swedish (see Figure 7). The HTML highlighting for parts of the text was also specified using the import functionality.

4. Acknowledgements

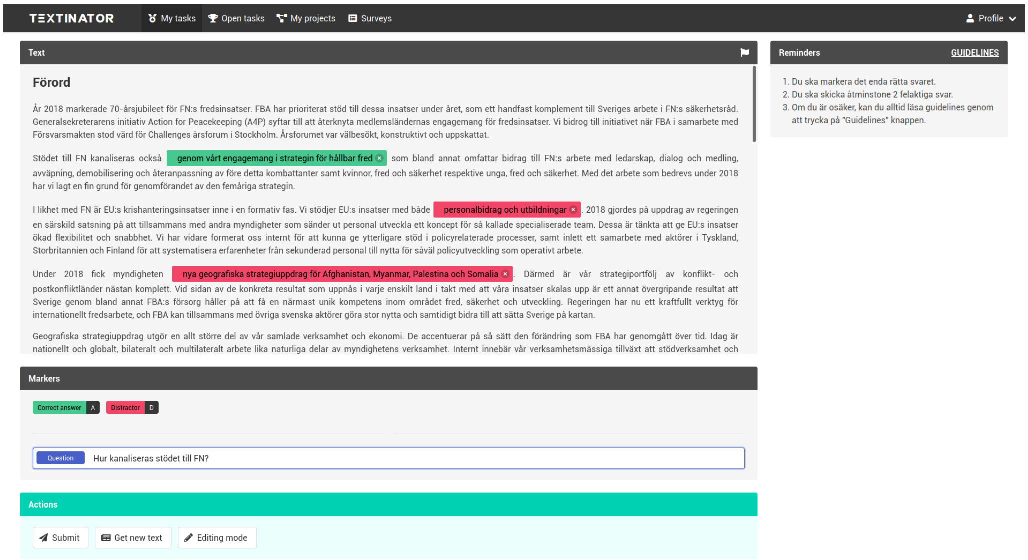
This work was supported by Vinnova (Sweden's Innovation Agency) within project 2019-02997.

[2] https://www.google.com/forms/about/
[3] https://www.surveymonkey.com/
[4] https://www.qualtrics.com
[5] https://bit.ly/3fuPL3V
[6] https://bit.ly/327pYf3



Larmet om branden kom klockan 22.31.– När 1 vi × kom fram kunde 1 vi × konstatera att branden var fullt utvecklad och hade spridit sig till bilar och en husbil i närheten, säger Daniel Nordin, operativ chef vid 1 Södertörns brandförsvar ×.Räddningstjänsten hade ingen möjlighet att gå in och rökdyka.– Eftersom villan var helt övertänd har 1 vi × inte kunnat gå in. Insatsen fokuserades i stället på att förhindra ytterligare spridning.Polisen har rutinmässigt inlett en förundersökning om misstänkt mordband. Polisens tekniker kan påbörja sitt × arbete först när brandhärden svalnat.Räddningstjänsten lät under morgonen villan brinna ner under kontrollerade former.– 1 Vi × är där och övervakar så att elden inte sticker iväg någonstans. Det × kommer att brinna flera timmar till, sa tjänstgörande larmoperatör vid räddningstjänsten tidigt på fredagsmorgonen.Som mest har ett 15-tal brandmän deltagit i insatsen.Vid 10-tiden på morgonen uppger polisen att branden ännu inte är helt släckt, så polisens tekniker har inte kunnat undersöka huset. – Deras × arbete påbörjas ordentligt under dagen. Man kan inte säga säkert om det × fanns någon där inne eller inte, säger Anders Bryngelsson, polisens presstalesperson.Finns det × indikationer på det ×?– Inte vad jag × känner till, men 1 vi × måste kunna utesluta det × så att säga.

Figure 2: Textinator’s UI for pronoun resolution



Förord

År 2018 markerade 70-årsjubileet för FN:s fredsinsatser. FBA har prioriterat stöd till dessa insatser under året, som ett handfast komplement till Sveriges arbete i FN:s säkerhetsråd. Generalsekreterarens initiativ Action for Peacekeeping (A4P) syftar till att återknyta medlemsländernas engagemang för fredsinsatser. Vi bidrog till initiativet när FBA i samarbete med Försvarsmakten stod värd för Challenges årsforum i Stockholm. Årsforumet var välbesökt, konstruktivt och uppskattat.

Stödet till FN kanaliseras också genom vårt engagemang i strategin för hållbar fred × som bland annat omfattar bidrag till FN:s arbete med ledarskap, dialog och medling, avväpning, demobilisering och återanpassning av före detta kombattanter samt kvinnor, fred och säkerhet respektive unga, fred och säkerhet. Med det arbete som bedrevs under 2018 har vi lagt en fin grund för genomförandet av den femåriga strategin.

I likhet med FN är EU:s krishanteringsinsatser inne i en formativ fas. Vi stödjer EU:s insatser med både personalbidrag och utbildningar ×. 2018 gjordes på uppdrag av regeringen en särskild satsning på att tillsammans med andra myndigheter som sänder ut personal utveckla ett koncept för så kallade specialiserade team. Dessa är tänkta att ge EU:s insatser ökad flexibilitet och snabbhet. Vi har vidare formerat oss internt för att kunna ge ytterligare stöd i policyrelaterade processer, samt inlett ett samarbete med aktörer i Tyskland, Storbritannien och Finland för att systematisera erfarenheter från sekunderad personal till nytta för såväl policyutveckling som operativt arbete.

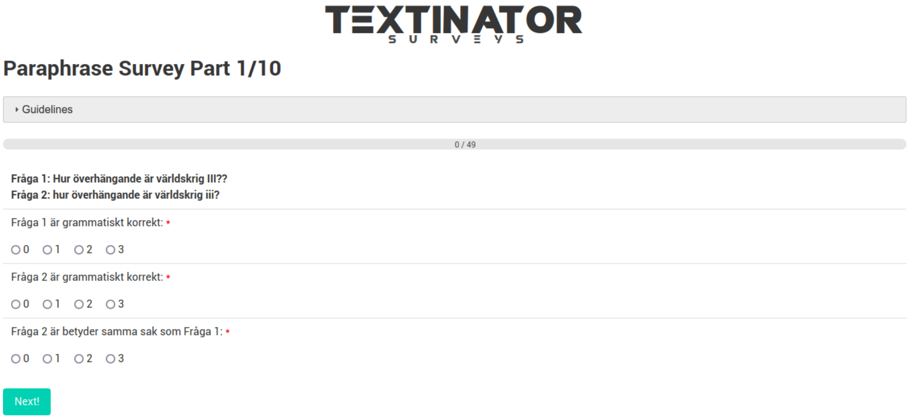
Under 2018 fick myndigheten nya geografiska strategiuppdrag för Afghanistan, Myanmar, Palestina och Somalia ×. Därmed är vår strategiportfölj av konflikt- och postkonfliktländer nästan komplett. Vid sidan av de konkreta resultat som uppnås i varje enskilt land i takt med att våra insatser skalas upp är ett annat övergripande resultat att Sverige genom bland annat FBA:s försorg håller på att få en närmast unik kompetens inom området fred, säkerhet och utveckling. Regeringen har nu ett kraftfullt verktyg för internationellt fredsarbete, och FBA kan tillsammans med övriga svenska aktörer göra stor nytta och samtidigt bidra till att sätta Sverige på kartan.

Geografiska strategiuppdrag utgör en allt större del av vår samlade verksamhet och ekonomi. De accentuerar på så sätt den förändring som FBA har genomgått över tid. Idag är nationellt och globalt, bilateralt och multilateralt arbete lika naturliga delar av myndighetens verksamhet. Internt innebär vår verksamhetsmässiga tillväxt att stödverksamhet och

1. Du ska markera det enda rätta svaret.
2. Du ska skicka åtminstone 2 felaktiga svar.
3. Om du är osäker, kan du alltid läsa guidelines genom att trycka på "Guidelines" knappen.

Question: Hur kanaliseras stödet till FN?

Figure 3: Textinator’s UI for multiple choice question answering task



Paraphrase Survey Part 1/10

Fråga 1: Hur överhängande är världskrig III??
Fråga 2: hur överhängande är världskrig iii?

Fråga 1 är grammatiskt korrekt: *

○ 0 ○ 1 ○ 2 ○ 3

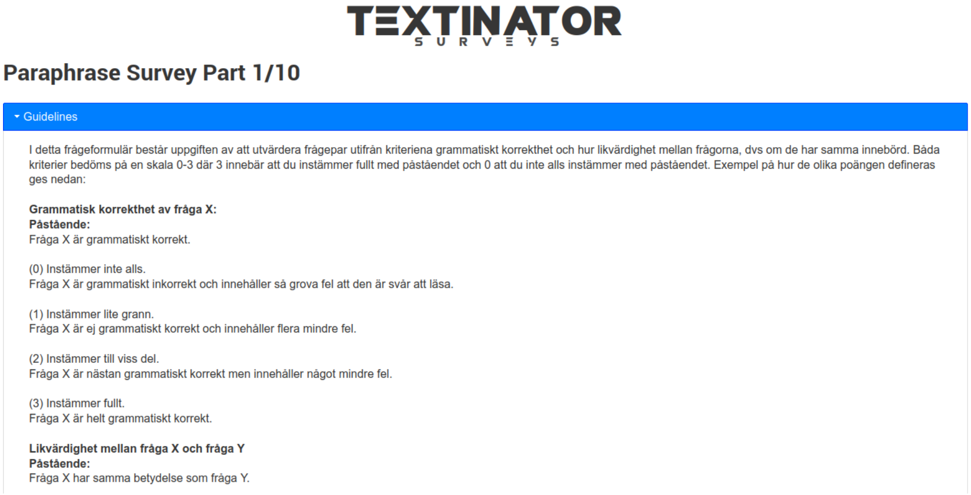
Fråga 2 är grammatiskt korrekt: *

○ 0 ○ 1 ○ 2 ○ 3

Fråga 2 är betyder samma sak som Fråga 1: *

○ 0 ○ 1 ○ 2 ○ 3

Figure 4: Textinator’s UI for paraphrase quality evaluation used



TEXTINATOR SURVEYS

Paraphrase Survey Part 1/10

▾ Guidelines

I detta frågeformulär består uppgiften av att utvärdera frågepar utifrån kriteriena grammatiskt korrekthet och hur likvärdighet mellan frågorna, dvs om de har samma innebörd. Båda kriterier bedöms på en skala 0-3 där 3 innebär att du instämmer fullt med påståendet och 0 att du inte alls instämmer med påståendet. Exempel på hur de olika poängen defineras ges nedan:

Grammatisk korrekthet av fråga X:
Påstående:
Fråga X är grammatiskt korrekt.

(0) Instämmer inte alls.
Fråga X är grammatiskt inkorrekt och innehåller så grova fel att den är svår att läsa.

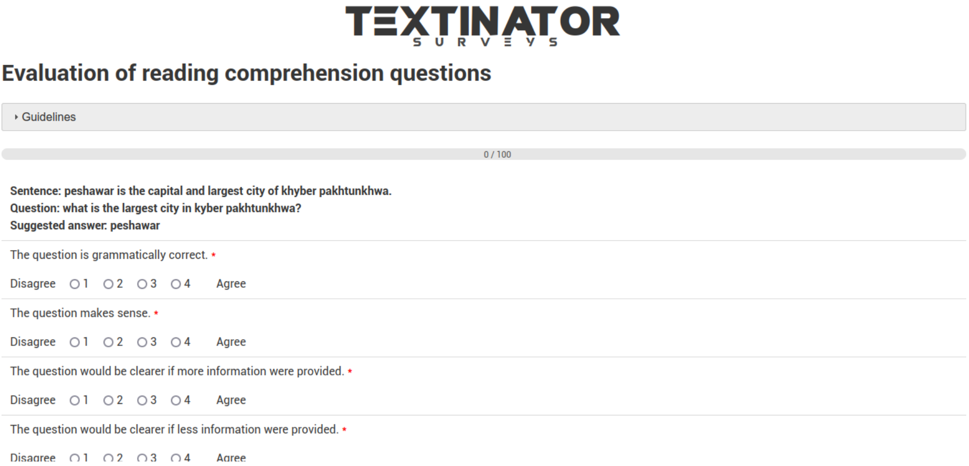
(1) Instämmer lite grann.
Fråga X är ej grammatiskt korrekt och innehåller flera mindre fel.

(2) Instämmer till viss del.
Fråga X är nästan grammatiskt korrekt men innehåller något mindre fel.

(3) Instämmer fullt.
Fråga X är helt grammatiskt korrekt.

Likvärdighet mellan fråga X och fråga Y
Påstående:
Fråga X har samma betydelse som fråga Y.

Figure 5: Textinator's UI for evaluation guidelines



TEXTINATOR SURVEYS

Evaluation of reading comprehension questions

▸ Guidelines

0 / 100

Sentence: peshawar is the capital and largest city of khyber pakhtunkhwa.
Question: what is the largest city in kyber pakhtunkhwa?
Suggested answer: peshawar

The question is grammatically correct. *

Disagree ○ 1 ○ 2 ○ 3 ○ 4 Agree

The question makes sense. *

Disagree ○ 1 ○ 2 ○ 3 ○ 4 Agree

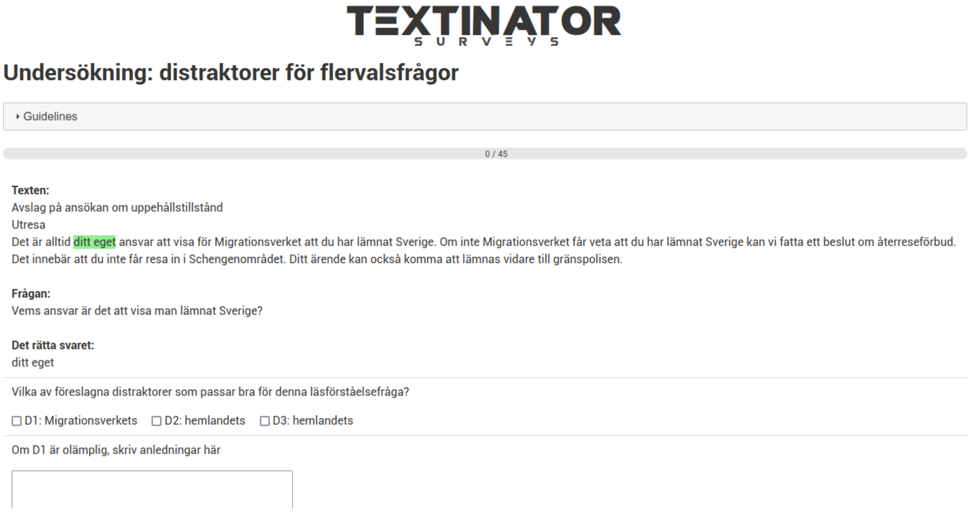
The question would be clearer if more information were provided. *

Disagree ○ 1 ○ 2 ○ 3 ○ 4 Agree

The question would be clearer if less information were provided. *

Disagree ○ 1 ○ 2 ○ 3 ○ 4 Agree

Figure 6: Textinator's UI for question-answer pair quality evaluation



TEXTINATOR SURVEYS

Undersökning: distraktorer för flervalsfrågor

▸ Guidelines

0 / 45

Texten:
Avslag på ansökan om uppehållstillstånd
Utresa
Det är alltid ditt eget ansvar att visa för Migrationsverket att du har lämnat Sverige. Om inte Migrationsverket får veta att du har lämnat Sverige kan vi fatta ett beslut om återreseförbud. Det innebär att du inte får resa in i Schengenområdet. Ditt ärende kan också komma att lämnas vidare till gränspolisen.

Frågan:
Vems ansvar är det att visa man lämnat Sverige?

Det rätta svaret:
ditt eget

Vilka av föreslagna distraktorer som passar bra för denna läsförståelsefråga?

□ D1: Migrationsverkets □ D2: hemlandets □ D3: hemlandets

Om D1 är olämplig, skriv anledningar här

Figure 7: Textinator's UI for distractor quality evaluation form

5. Bibliographical References

Ahlenius, C. (2020). Automatic pronoun resolution for Swedish. Available at http://kth.diva-portal.org/smash/record.jsf?pid=diva2:1520819.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

Daudert, T. (2020). A web-based collaborative annotation and consolidation tool. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7053–7059.

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., and Choi, Y. (2021). Scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294*.

Kalpakchi, D. and Boye, J. (2021a). BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset. *arXiv preprint arXiv:2108.03973*.

Kalpakchi, D. and Boye, J. (2021b). Quinductor: a multilingual data-driven method for generating reading-comprehension questions using Universal Dependencies. *arXiv preprint arXiv:2103.10121*.

Kirk, H. R., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., Asano, Y., et al. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing Systems*, 34.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (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.

Lindqvist, N. (2021). Automatic question paraphrasing in Swedish with deep generative models. Available at http://kth.diva-portal.org/smash/record.jsf?pid=diva2:1554622.

McCormick, C. (2013). Countries with better English have better economies. Available at https://hbr.org/2013/11/countries-with-better-english-have-better-economies.

Morton, T. S. and LaCivita, J. (2003). WordFreak: an open tool for linguistic annotation. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Demonstrations*, pages 17–18.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Tengvall, T. (2020). A method for automatic question answering in Swedish based on BERT. Available at http://kth.diva-portal.org/smash/record.jsf?pid=diva2:1501493.

Tkachenko, M., Malyuk, M., Shevchenko, N., Holmanyuk, A., and Liubimov, N. (2020). Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

Appendix: Details of the Comparison of Annotation Tools

In this appendix we provide detailed explanation for the scores assigned in Table 1. For typographic purposes, we have assigned each annotation tool under comparison a unique identifier as follows.

T1	WordFreak (Morton and LaCivita, 2003)
T2	MMAX2 (Müller and Strube, 2006)
T3	BRAT (Stenetorp et al., 2012)
T4	WebAnno (Yimam et al., 2013)
T5	GATE Teamware (Bontcheva et al., 2013)
T6	INCEpTION (Klie et al., 2018)
T7	AWOCATo (Daudert, 2020)
T8	*Doccano* (Nakayama et al., 2018)
T9	*Label Studio* (Tkachenko et al., 2020)
T10	`Textinator` (ours)

Please see Tables 2 to 9 on the next pages.

Tool	Task types
T1	trees (constituency and dependency parsing), span markers (named entity recognition), relations (coreferences)
T2	span markers (cfg), relations (cfg), trees (dependency)
T3	relations (cfg), trees (dependency), span markers (cfg)
T4	span-based (cfg), relations (cfg), trees (dependency)
T5	span-based (cfg), relations (cfg)
T6	span-based (cfg), relations (cfg), trees (dependency), free-text information (via document metadata)
T7	real-valued numbers (sentiment annotation), text markers (binary or tertiary text classification), free-text information (cfg)
T8	span markers (cfg), text markers (cfg), free-text information (machine translation)
T9	span markers (named entity recognition), text markers (cfg), relations (cfg)
T10	span markers (cfg), relations (cfg), marker groups (cfg), ranked marker groups (cfg), text markers (cfg), real-valued numbers (cfg), free-text information (cfg)

Table 2: Detailed comparison results for the task coverage axis (A1). “cfg” stands for “configurable” and means that many annotation tasks of this type could potentially be created

	T1	T2	T3	T4	T5	T6	T7	*T8*	*T9*	T10
C1: static UI elements	✗	✗	✗	✗	✗	✗	✗	✔	✗	✔
C2: static markers	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔
C2: static relations	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔
C4: dynamic markers	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔
C4: dynamic relations	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔

Table 3: Detailed comparison results for the internationalization axis (A2)

	T1	T2	T3	T4	T5	T6	T7	*T8*	*T9*	T10
C1: Custom markers for spans of text	✗	⊘	⊘	👁	👁	👁	⊘	👁	⊘	👁
C2: Custom appearance of the markers	✗	⊘	⊘	✗	👁	👁	⊘	👁	⊘	👁
C3: Custom hotkeys for markers?	✗	✗	⊘	✗	✗	👁	✗	👁	⊘	👁
C4: Custom information associated with any of the markers	✗	✗	👁	👁	👁	👁	✗	✗	👁	👁
C5: Custom relations for the markers?	✗	⊘	⊘	👁	✗	👁	✗	✗	⊘	👁
C6: Custom constraints on the relations	✗	⊘	⊘	👁	✗	👁	✗	✗	✗	👁
C7: Custom appearance of relations	✗	⊘	⊘	✗	✗	👁	✗	✗	⊘	👁
C8: Custom hotkeys for relations	✗	✗	⊘	✗	✗	✗	✗	✗	✗	👁
C9: Custom marker groups to be annotated as a unit	✗	✗	✗	✗	✗	✗	✗	✗	✗	👁
C10: Custom appearance of marker groups	✗	✗	✗	✗	✗	✗	✗	✗	✗	👁
C11: Custom constraints on the marker groups	✗	✗	✗	✗	✗	✗	✗	✗	✗	👁
C12: Custom UI layout (to any degree)	✗	✗	✗	👁	✗	👁	⊘	✗	⊘	✗

Table 4: Detailed comparison results for the customization axis (A3). 👁 denotes that the objective can be achieved via UI interactions, ⊘ denotes that it can be achieved only by means other than UI interactions, but requires no programming skills, ✗ denotes that it cannot be achieved or programming skills are required

Tool	Features
T1	found none
T2	found none
T3	an address for each annotation; detailed annotation process measurement
T4	project progress as an annotator-document matrix; progress of individual annotator; dataset completion statistics.
T5	annotation status monitoring (displayed as pie charts); tracking average annotation time; per-annotator statistics; per-document statistics
T6	an overview which documents have already been annotated and who annotated them; statistics about the annotated tokens and sentences (using MTAS); progress of individual annotator;
T7	found none
T8	found none
T9	found none in the community edition
`T10`	dataset completion statistics; user progress tracking; distribution of the annotation lengths; flagging problems with text

Table 5: Detailed comparison results for the annotation progress tracking axis (A4)

Tool	Features
T1	found none
T2	list-based visualization and highlighting of differences; Kappa statistic for nominal attributes; calculating inter-annotator agreement for relations
T3	integrated annotation comparison
T4	the curator can open and compare annotations made by multiple annotators; the curator can reconcile annotations; can calculate Kappa and Tau measures
T5	calculates IAA metrics (including f-measure and Kappa); a visual annotation comparison tool; adjudication editor to reconcile annotations
T6	calculates IAA metrics (Cohen's kappa, Fleiss' kappa, Krippendorff's alpha); UI for merging annotations; handling abandoned documents by reassining annotators after time-out.
T7	automatic consolidation if the number of required annotations per text, and the stipulated standard deviation are defined; manual consolidation
T8	found none
T9	found none in the community edition
`T10`	none

Table 6: Detailed comparison results for the quality assurance axis (A5)

Tool	Features
T1	found none
T2	found none
T3	found none
T4	separate roles for annotator, curator and project manager;
T5	out-of-the-box roles of administrator, project manager and annotator; the possibility to define custom roles; defining number of annotators per document; benchmarking annotator's performance against gold data; defining workflows as custom graph-based schemas
T6	separate roles for annotator, administrator, project creator, or remote API access; dynamic assignment if you want each document to be annotated by a certain number of annotators; project versioning; guest annotators; invite links
T7	found none
T8	a separate role for annotator
T9	found none in the community edition
`T10`	separate roles for annotator, project manager, translator and system administrator; custom roles can be defined; defining whether each document will be annotated once or by each annotator; monthly time reports per annotator

Table 7: Detailed comparison results for the administration quality axis (A6)

	T1	T2	T3	T4	T5	T6	T7	*T8*	*T9*	`T10`
C1: text-form tutorials	✘	✔	✔	✔	✔	✔	✔	✔	✔	✔
C2: video tutorials	✘	✘	✘	✔	✘	✔	✘	✘	✔	✔
C3: sandbox demo	✔	✔	✔	✘	✘	✔	✘	✔	✔	✘
C4: source code documentation	✔	✘	✘	✔	✘	✔	✘	✘	✔	✔

Table 8: Detailed comparison results for the documentation quality axis (A7)

	T1	T2	T3	T4	T5	T6	T7	*T8*	*T9*	`T10`
C1: out of the box suggestions on the fly	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘
C2: external models for on-the-fly suggestions	✘	✘	✘	✘	✘	✔	✘	✘	✔	✘
C3: pre-annotating texts	✔	✘	✔	✘	✔	✘	✘	✘	✔	✔
C4: training custom suggestion models	✘	✘	✘	✔	✘	✔	✘	✘	✔	✘
C5: configurable sanity checks	✘	✘	✔	✔	✘	✔	✘	✘	✘	✔
C6: active learning	✘	✘	✘	✘	✘	✔	✘	✘	✔	✘
C7: spelling correction	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
C8: grammar correction	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘

Table 9: Detailed comparison results for the intelligent assistance axis (A8)