Is NLP Ready for Standardization?

Lauriane Aufrant
Inria, Rocquencourt, France
first.last@inria.fr

Abstract
While standardization is a well-established activity in other scientific fields such as telecommunications, networks or multimedia, in the field of AI and more specifically NLP it is still at its dawn. In this paper, we explore how various aspects of NLP (evaluation, data, tasks...) lack standards and how that can impact science, but also the society, the industry, and regulations. We argue that the numerous initiatives to rationalize the field and establish good practices are only the first step, and developing formal standards remains needed to bring further clarity to NLP research and industry, at a time where this community faces various crises regarding ethics or reproducibility. We thus encourage NLP researchers to contribute to existing and upcoming standardization projects, so that they can express their needs and concerns, while sharing their expertise.

1 Introduction
Most of the Natural Language Processing community remains estranged from standardization. As this is already common practice in many computer science fields, including telecommunications, networks and multimedia, what is making NLP so special in that regard? Zielke (2020) has already asked the question of the potential barriers and benefits of standardization work in the broader field of Artificial Intelligence, which is now becoming a reality. Here we propose to deepen that discussion by investigating the more specific context of NLP and its standardization needs.

Standards are normative documents produced by Standards Developing Organizations (SDOs) such as ISO. In practice, they can be of various nature and content. Some of them are terminological references that establish shared terms and definitions for a technical domain. For instance, the ISO Online Browsing Platform\(^1\) indexes all existing ISO definitions. Other standards rather describe a reference framework, which can prove useful for bootstrapping new activities or rationalizing existing ones. Standards can also provide technical specifications for data, systems or procedures. This notably includes quality specifications, as well as interoperability ones (APIs, protocols, etc.). For instance, the C language (ISO/IEC 9899), the MP3 coding (ISO/IEC 11172-3), the Latin-1 charset (ISO/IEC 8859-1) and the 2-letter language codes (ISO 639-1) are all examples of standards. Technical specifications are often associated with certification, and test suites can be developed to assess compliance with standards offering sufficient formalism (e.g. syntax checkers, protocol testing, or dedicated measurements against standard thresholds).

Standards are written by volunteer experts from various backgrounds (scientific, legal, standardization experts, etc.). In most SDOs, registration is open to anyone willing to contribute, usually through a mirror committee within their national standards organization. Experts collaborate within working groups and decisions are taken by consensus\(^2\) – across countries, but also across backgrounds, across sectors, across technical fields. This approach makes standardization a rather slow process (with up to 3 years to establish some standards), but it also ensures the strength of the agreements. Most SDOs plan a mandatory revision of published standards every few years, without which they are declared obsolete and withdrawn, in order to ensure that standards remain up to date with technological and societal evolutions.

Standards are especially important for the indus-

\(^1\)https://www.iso.org/obp

\(^2\)Compared to unanimity, where everyone supports the decision, consensus means that noone objects to the decision. This decision process helps to identify middle ground solutions that everyone in an heterogeneous group can find acceptable, whereas an unanimity requirement would bear the risk of freezing projects due to unsolvable cultural differences or diverging interests.
try and for regulatory authorities – but they can apply on very technical fields, including scientific topics, and therefore affect the research community as well. This paper investigates how NLP standardization could impact our community, offering both challenges and opportunities.

After a brief review of the current state of NLP standardization initiatives (§2), we explore standardization gaps pertaining to NLP evaluation (§3), data and formats (§4), tasks (§5) and higher-level concepts (§6). Having built a broader view of how NLP standardization could benefit research, but also the society, industry and regulations (§7), we then conclude on possible contributions that NLP researchers could add to those efforts (§8).

2 Existing initiatives towards NLP standardization

Within the NLP community, most of the standardized material is actually de facto standards: data, tools or methodology that are consensually used throughout the field, even though they don’t have any official status and have not necessarily gone through the formalization process that official standards offer. Such de facto standards often result from past shared tasks: for instance, the mteval-v13a.pl evaluation script from the WMT shared tasks series has been used for years as the reference evaluation script by a large part of machine translation research. Similarly, the CoNLL-X format for dependency treebanks has been widely adopted following the corresponding CoNLL shared task (Buchholz and Marsi, 2006). As for event detection, the definition of the task itself is fully driven by the ACE 2005 campaign (Walker et al., 2006), to the point that it is sometimes referred to as “ACE event detection” (Chen et al., 2018). Recent years have seen however the growth of the Universal Dependencies initiative (Nivre et al., 2016), for establishing common guidelines for treebank annotation across languages; considering the breadth of its contributors and its sustained efforts for guidelines formalization, this project has now become very close to standardization work and could be considered as an SDO.

Regarding official standardization initiatives relevant for NLP, the most established one is ISO’s Technical Committee 37 (Language and Terminology), and more prominently its subcommittees 4 (Language resource management, created in 2001)\(^3\) and 5 (Translation, interpreting and related technology, created in 2012). With a strong focus on corpora and annotation, these groups have notably released a number of annotation framework standards (e.g. for TIGER-XML or TEI), which are extensively used by the corresponding industry; they also co-organize with ACL the ISA workshop series on Interoperable Semantic Annotation (Bunt, 2021). Yet this focus on data leaves the algorithmic and evaluation parts of NLP largely unaddressed.

ISO-IEC’s Joint Technical Committee 1 (Information Technology) has created in 2017 its subcommittee 42 on Artificial Intelligence. This one is more concerned with algorithms, development methodology, and system evaluation, but at the higher level of AI in general and not delving into NLP-specific aspects. To date, its NLP-related activity has focused on defining a few major concepts, such as NLG, question answering, or OCR. Concurrently, other global SDOs such as ITU-T have also explored NLP standardization, but mostly in the context of specific use cases (e.g. ITU-T H.862.5 “Emotion enabled multimodal user interface based on artificial neural networks”) rather than the NLP field in general. Hence their remains a gap in terms of NLP standardization.

The European counterparts of ISO-IEC (CEN-CENELEC) have thus created in 2021 their own Joint Technical Committee 21 on Artificial Intelligence, with a group dedicated to kickstarting activities on speech and NLP (ad-hoc group 4, AI systems for human language processing, on track to become persistent in 2022). This is the first page of a new chapter, roadmaps are being written today. Now is thus the right time to question what can, and should, be standardized within our field.

3 Is NLP evaluation ready for standardization?

The reproducibility crisis that has spread through the field in recent years (Belz et al., 2021b; Lucic et al., 2022) has renewed the community’s interest for fair and reliable evaluation. This has led in the 2020s to a blooming of workshops and shared tasks dedicated to evaluation means on speech and NLP (ad-hoc group 4, Benchmarking: Past, Present and Future (Eger et al., 2020; Belz et al., 2021a; Bosselut et al., 2021; Church

\(^3\)See (Romary, 2015) for a detailed introduction to subcommittee 4’s activities.
et al., 2021). But those concerns are not new, as illustrated by the 4REAL workshops organized in the 2010s (Branco et al., 2016). Even the terminology of reproducibility has been the topic of debate and clarification attempts for many years in the machine learning community (Drummond, 2009). The existence of the LREC conference series itself is a token of that interest for standardized evaluation, with its first edition dedicating a whole workshop to the lack of a shared strategy, definition and infrastructure for system evaluation (ELSE, 1998; McTait and Choukri, 2003).

For domains like human evaluation of NLG, the need for more consistent practices is clear enough, and Howcroft et al. (2020) have already advocated for producing standards on both the methodology and the terminology, based on their review of 20 years of NLG with conflicting evaluation criteria. Yet here we argue that even cases with seemingly straightforward automated evaluation can suffer today from the lack of standards.

### 3.1 On defining metrics

One of the challenges of standardized evaluation is to ensure that the metrics used are defined in a way that leaves no place to ambiguity, which in practice is rarely the case in the field. For instance, even the well-known F1-score, despite its very formal definition as the harmonic mean of precision \( \left( \frac{TP}{TP + FP} \right) \) and recall \( \left( \frac{TP}{TP + FN} \right) \), becomes ambiguous when applied to tasks such as Named Entity Recognition.

A first issue resides in the common practice of casting this chunking task into a sequence labelling task, through BIO-style token-level encoding: B-... labels denote the first token of an entity, I-... labels other tokens within the entity, and 0 labels other tokens. This raises the question of which objects are considered for true/false positives/negatives: those labels, or the chunks. For instance, with B-PER I-PER 0 B-LOC 0 as a reference sequence, predicting B-PER 0 0 B-LOC 0 yields a score of 60 (micro-/macro) while if evaluated as a sequence labeling task by looking at tokens (67 F1 if 0 is not considered a class), but 0 F1 if evaluated as a chunking task by looking at the predicted chunks (here with exact match). While most experienced researchers know to prefer the latter (and know where to get that information), the youngest researchers as well as industry practitioners are not necessarily aware of that implicit rule. Such confusion can in turn lead to incorrect comparison of models, or incorrect reporting of product performance.

Taillé et al. (2020) report other underspecified aspects, such as the criteria to accept true positives (with or without typing, with partial or exact match...), the use of micro- or macro-averaging, or the existing practice to ignore some classes (such as Other or MISC). As they highlight, these issues also propagate to evaluation of relation extraction, and just one of those can already lead to overestimating the results by up to +3 F1 on a widely used dataset.

### 3.2 On implementing metrics

Another challenge is the underspecification of implementation details for those evaluation metrics, even when the metric itself has a non-ambiguous definition.

For machine translation, Post (2018) investigates the divergence in scores that can result from different implementations of the BLEU metric, based on diverging choices of parameters and preprocessing (e.g. the maximum n-gram length, the number of references, or user-supplied and/or metric-internal tokenization). He reports up to 1.8 BLEU difference when varying only the tokenization used for scoring, which is actually more than the gains measured for BPE (Sennrich et al., 2016), which was a game changer for neural machine translation.

Such variations in implementation can occur even in cases as seemingly simple as using F1-scores for classification: for instance Belz (2021) compares concurrent reproduction studies of the same text classifier, and reports score divergences up to 5.2 F1 due to metric reimplementation.

Another source of implementation divergence is the procedure adopted to deal with invalid outputs (ill-formatted, impossible sequences, etc.). In the case of Named Entity Recognition, Lignos and Kamya (2020) investigate how different strategies to repair invalid BIO sequences within the scorer can impact the measured F1, a condition which according to Palen-Michel et al. (2021) also affects the gold labels in a number of renowned datasets, and leads to differences up to 3.25 F1 in a realistic scenario. For the BIOES encoding scheme alone (one of BIO’s competitors, see §4),
Kroutikov (2019) numbers at least 7776 different strategies that could be adopted to repair invalid label pairs.

3.3 Tooling to the rescue?

As a means to circumvent those pitfalls, Lignos and Kamyab (2020) advocate for never reimplementing evaluation metrics and relying instead on third-party reference tools. This is in line with Post (2018)’s strategy to release the SacreBLEU package, with the hope that its configurability, documentation, ease of use and variant reporting will enable standardized evaluation. Can tools alone indeed fill in for standards?

The main issue with that view is that it supposes that tools are correctly used. However, Marie et al. (2021) unveil that the growing number of users of SacreBLEU are in practice often misusing it (not reporting the variant used, comparing its scores with other scorers, etc.). Similarly, Palen-Michel et al. (2021) release SeqScore as a possible reference tool for named entity recognition evaluation, but they do so based on the failure of previous de facto standard tools. For instance, Akbik et al. (2019) observe that their previous paper (which has now over 1000 citations) had overestimated its results by up to 0.8 F1, because they used the official CoNLL-03 evaluation script (designed for BIO) on a BIOES-encoded dataset. On a side note, it can also happen that the most popular scorer simply contains a bug – how can this be assessed if the tool itself serves as standard?

Another possible approach would be to rely more heavily on Kaggle-style benchmarking platforms that enable fairer comparison than standalone evaluation tools, by offering uniform and fully reproducible evaluation conditions. The issue here is that such practices can arbitrarily foster inadequate evaluation. Bowman and Dahl (2021) now consider NLU evaluation “broken” due to benchmark-driven standardization of practices: a number of those benchmarks are actually rewarding “unreliable and biased systems”. They leave no place to reflect upon a given system’s appropriate evaluation setting, and instead incentivize gaming the numbers. Church and Hestness (2019) review 25 years of evaluation practices and show how the rigour efforts that have led to such benchmarks are now pushing against their initial purpose of bringing more insights to “content-free debates”. Extensive reliance on benchmarking platforms for reproducible evaluation would only strengthen the reliance on benchmark data, hence those pitfalls. Massive use of identical data and data splits is itself an issue, as leading to community-wide overfitting to the test set (Gorman and Bedrick, 2019). Overall, leaderboards have drawn a lot of criticism in recent years (Rogers, 2019; Ethayarajh and Jurafsky, 2020; Kiela et al., 2021) and are therefore a poor candidate to address the lack of standards.

Instead of producing and relying on tools, typical standardization work would rather approach the issue by writing comprehensive specifications of the evaluation metrics (detailing their computation, their usage, their meaning), which can in turn apply on tools. This includes providing the means to verify that a given scorer or a given evaluation protocol is compliant with the specification. Hence comparable evaluation can be formally ensured, but not at the cost of insights and appropriateness.

3.4 Does it matter?

So the lack of standards leads to more imprecision in the measures and less rigorous comparisons. Is that really an issue, as long as those numbers are high and continue increasing, whatever the criteria? Haven’t experimental sciences handled imprecision for centuries, and accepted that challenge as part of the job?

According to Morey et al. (2017), such imprecision has already endangered scientific progress in whole fields of NLP: in their review of several years of contributions in discourse parsing, they discover that the various conclusions drawn on the benefits of distributed representations are mostly wrong in that field. What was considered a huge improvement, with 24 to 51% relative error reduction depending on the metric, was actually a gain of 11 and 16% for two of the metrics, and a loss of 15 and 53% for the other two. Here the culprit was the choice to macro-average over documents in some but not all of the works, following practices existing in different communities.

The lack of standards can thus lead to misinterpreting regress as progress. But it can also affect the wider world outside of research. For instance when a contract is signed, and B2B products are to be developed according to a given performance level specified in the contract, there should be no place to ambiguity. Who should be the judge of whether the contract is fulfilled, if the bar is met by one implementation variant of the metric, but not...
the other? And what if a regulation contains such performance requirement?

Comparability is also a strong enabler for individual rights as consumers. Potential users should be able to make an informed choice when comparing existing products. Transparency regulations can contribute to that, but that information becomes meaningless if the same number can be interpreted differently depending on implementation details.

3.5 Can standards hinder research?

Scientific concerns regarding NLP evaluation go in fact way beyond the need for fair comparison. A number of automated metrics in wide use today have poor correlation with human judgment, and a lot of research efforts have been devoted to designing more relevant metrics. Notoriously, WMT has been running an annual shared task on machine translation evaluation metrics since 2008 (Bojar et al., 2016; Mathur et al., 2020; Freitag et al., 2021), thereby consolidating the community consensus that the BLEU metric certainly has its utility, but also a number of shortcomings (Reiter, 2018), and it is far from being the best metric in existence. METEOR, chrF, CharacTer, BERTScore (Banerjee and Lavie, 2005; Popović, 2015; Wang et al., 2016; Zhang et al., 2019) are just a few examples among a broad panel of often more appropriate metrics, even though the single-best “one-size-fits-all” metric has not been found yet.

One possible fear with standardization could then be to prevent researchers from pursuing their quest for the best metric, or simply to prevent them from using in their work another good metric instead of BLEU – leading again to fostering bad evaluation practices and limiting the insights brought to future research. However, standards do not need to be compulsory. It is quite possible to write them in a way that preserves that research freedom, but still brings some order and clarity. BLEU is not the best metric, but BLEU is nevertheless preferable to exotic approaches such as measuring an F1-score at the sentence level (true positive if the sentence is an exact match). Are we confident that all practitioners that may have to evaluate a machine translation system at some point (including e.g. software developers in the industry) are aware of that? Can we at least give formal existence to that tiny piece of knowledge?

It is indeed a fact that in a number of cases in NLP evaluation, it is not necessarily known what is the most appropriate choice among the various existing variants. It can also be use case-dependent. And standards in such context are not meant to arbitrarily foster one option among the others. Their role here would rather be to formally reference and specify the existing relevant options (pushing away the ones that are already known to be inappropriate), and offer practical ways to declare, identify or verify which one of those options has indeed been used in a given paper or product.

4 Are NLP data and formats ready for standardization?

Yes they are, and they have been as early as 1993, when EAGLES (Expert Advisory Group for Language Engineering Standards) was established to develop such standards. Ide et al. (2017) review 30 years of community progress from confusion to de facto standards to standards. However, despite marked efforts from ISO’s Technical Committee 37, this paradigm has only been adopted so far in some parts of the field, and much progress remains to achieve for fully standardizing NLP annotations.

In particular, Ide et al. (2017) underlie the need to better standardize the content of annotations. While many (although not all) corpus authors have gone through the formalization process of writing annotation guidelines, this has mostly led to a profusion of co-existing guidelines for the same task. The case of dependency parsing is interesting in that regard, as the Universal Dependencies project managed to unify most of the pre-existing annotation schemes, while preserving their idiosyncrasies (Nivre et al., 2016). Yet this is a success story that most parts of the field have not had so far.

In addition, annotation processes should include some quality control mechanisms, such as measuring inter-annotator agreement (Hovy and Lavid, 2010). However, there is poor consensus on what would be a “good” agreement value for a given task, depending on its complexity and subjectivity (Artstein and Poesio, 2008; Mathet et al., 2012). Are we even sure that inter-annotator agreement is an appropriate quality control (Wong and Lee, 2013; Passonneau and Carpenter, 2014; Plank et al., 2014; Boguslav and Cohen, 2017; Basile et al., 2021)? In recent years, the growing reliance on crowdsourcing has only strengthened the challenges, hence the pressing need for standardizing practices (Sabou et al., 2014).

Standardization gaps do not concern solely the
semantics of the annotation, but also their format. Looking at machine translation, parallel corpus formats include SGML, (for which WMT maintains a `wrap-xml.perl` script to preserve compatibility with scoring scripts), XML (with XCES for sentence alignment), TMX, bitext (two files with corresponding line numbers), but also tabular formats with per-language columns separated by either tabs or other separators. The OpusTools converters (Aulamo et al., 2020) support only part of that spectrum. As for named entity recognition, co-existing encoding schemes include IO, IOB (aka IOB1), BIO (aka IOB2), BIOES (aka IOBES), BILOU (aka BILOU) and BMEOW (Palen-Michel et al., 2021). And there are others (as in Malik and Sarwar, 2016). One can always write converters, but this is tedious work, and prone to introducing discrepancies in case of invalid sequences (see §3.2). Third-party open source converters can help (Lester, 2020), yet they usually support only some of the encoding schemes. Formats can further differ when considering the file format: whereas CoNLL-2003 was distributed as tabular IOB (Tjong Kim Sang and De Meulder, 2003), spaCy relies on JSON BILUO. And this is only for sequence tag schemes, while MUC-6 uses SGML (Grishman and Sundheim, 1996) and WiNER-fr prefers an offset-based scheme to directly encode the spans (Dupont, 2019).

In terms of input and output formats, NLP tools can already rely on a number of extensible pipelines such as Stanford CoreNLP or spaCy (Manning et al., 2014; Honnibal and Montani, 2017), as well as abstraction frameworks such as AllenNLP or PyText (Gardner et al., 2018; Aly et al., 2018) – but this differs from actual APIs designed for interoperability among products. Today such interoperability is mostly fostered by infrastructure-based initiatives such as the Language Application Grid (Ide et al., 2016, 2015). The European Language Grid project (Rehm et al., 2020a, 2021) now proposes to build an umbrella platform that hosts resources but also unifies NLP APIs through its “functional services” infrastructure. In addition, Kim et al. (2020) propose to standardize a web protocol for NLPaaS, while Rehm et al. (2020b) set a roadmap of interoperability levels to enable cross-platform workflows. Instead of duplicating those projects, the role of SDOs here would rather be to build upon those APIs, by escalating them into official standards with formal specifications.

Finally, data warrants data documentation. This is another area where individual initiatives have produced valuable guidelines on necessary metadata (Bender and Friedman, 2018). But work still remains to give that material more formalism and ensure consensus across communities.

### 5 Are NLP tasks ready for standardization?

Getting to the core of NLP, even the tasks themselves warrant further consideration for standardization. Indeed, NLP research has recently gained awareness that making further progress on NLU tasks now meant taking some detours to better define terms like “meaning” (Bender and Koller, 2020), “comprehension” (Dunietz et al., 2020) and the associated tasks. Yet even the basic expectations on inputs/outputs can be underspecified for some tasks. For instance, question answering can refer to various concrete tasks, such as multiple-choice answer selection (Aydin et al., 2014), span extraction (with or without paragraph retrieval) in the SQuAD style (Rajpurkar et al., 2016), free-form answering that can include multi-hop questions (Chen et al., 2019), or answering questions over knowledge bases (Fu et al., 2020), which don’t warrant the same algorithmic approaches. Gardner et al. (2019) propose to solve the conundrum by considering question answering as a format and splitting it from the definition of the task; yet even then the taxonomy remains dense (Rogers et al., 2021).

Information extraction is another field where tasks and their terminology are largely ill-defined. Even its primary task, named entity recognition, has been subject to a number of conceptual debates (Marrero et al., 2013). Entity linking is better delineated, but has been associated with a number of different names: entity linking, named entity linking, named entity disambiguation, named entity normalization... Are all of those terms synonymous, or do they slightly differ in scope? The literature has already proposed many definitions for entity linking, often inconsistently: for instance...
Shen et al. (2014) write both “Entity linking is the task to link entity mentions in text with their corresponding entities in a knowledge base” and “to link named entity mentions appearing in web text with their corresponding entities in a knowledge base, which is called entity linking”. This notably raises doubts as to whether entity linking applies only to named entities, or also to non-named entities (Paris and Suchanek, 2021). Or to non-named mentions of entities that have names? In the lack of terminological standards, presumably the best definition of the task is to look at how the corpus at hand has been annotated; but then the task definition can vary a lot from one dataset to another, so that evaluating an entity linking approach on multiple datasets may not make actual sense. Many other discrepancies could be listed here (e.g. relation extraction referring to either relation clustering, open information extraction, or relation classification), but in the end, the name “information extraction” itself is an ill-defined term, with a functional scope that varies a lot depending on individuals. So if a system is branded as an information extraction system, what are its functionalities supposed to be?

Time is not innocent in those terminological conflicts. Language modeling is one striking example of terminological drift. Historically, language models meant “a probability distribution over all possible word strings in a language” (Arisoy et al., 2012) – or even a next-word predictor, as in the n-gram paradigm: “language modeling, the problem of predicting the next word based on words already seen before” (Xu and Jelinek, 2004). But since 2018 and the advent of masked language models, the term “language model” has now shifted to refer to Transformer-based contextualized embeddings, regardless of any probability distribution, and not necessarily autoregressive (as in Ettinger, 2020).

Are these discrepancies an issue? Semantic drift is a natural phenomenon in any language, and a profusion of definitions also means a profusion of problems addressed by the community as a whole. However, trouble arises when using those task definitions to catalog or to assess existing systems: how to decide whether a given system meets one’s expectations, if it is branded with ambiguous functionalities? Achieving clarity on product capabilities is a matter of commercial interest for companies, and of consumer rights for individuals. But it can also affect scientific processes, as exemplified by the Great Misalignment Problem (Hämäläinen and Alnajjar, 2021) between blurry objectives, the actual task fulfilled by the system, and the task against which human evaluation is performed.

6 Are NLP concepts ready for standardization?

At a higher level, a number of concepts would also benefit from formal standards. This notably concerns the term “multilingual”, which has been used to describe very different properties, such as: a system that juxtaposes models for multiple languages (Otero and González, 2012), with or without internal language identification; an algorithm that does not rely on language-specific features or knowledge, and can therefore be trained on a dataset from any language (Johansson and Nugues, 2006; Szarvas et al., 2006), even though this does not guarantee actual language independence (Bender, 2011); or a single model that can indiscriminately process contents from many languages (Pires et al., 2019). Focusing only on the latter definition, how many is many? And how diverse? Can an Indo-European-only system be considered multilingual? In light of rising initiatives for fostering more language diversity in NLP research (Bender, 2019; Joshi et al., 2020), including a dedicated theme track at ACL 2022, it now appears pressing to establish consensual criteria on what renders a given system multilingual. Otherwise, how can progress in that matter be quantified?

Trustworthiness is another relevant concept for NLP systems, especially from the viewpoint of policy makers. The High-Level Expert Group on AI (2019) has notoriously established a list of AI trustworthiness characteristics, but they still lack shared actionable definitions. The concept of bias for instance, while subject to a growing interest in NLP research, is rarely formally defined in that literature, or with diverging senses (Blodgett et al., 2020), even though that conceptualization should be a prerequisite before defining the corresponding bias measures (Dev et al., 2021). “Robustness” is similarly overloaded, with meanings ranging from maintained performance on out-of-domain data (Bernier-Colborne and Langlais, 2020), on transformed data (Sanchez et al., 2018; Gan and Ng, 2019), or in presence of natural noise (Zhou et al., 2019), to specific defenses against adversarial attacks (Hsieh et al., 2019). A fortiori, there is no formal taxonomy on what kind of noise a “robust” NLP system should minimally handle: typos
only, or L2 learners grammar errors, lexical borrowings? Broken encoding? Or others? Those topics are at the heart of recent debates on the merits of biased splits in place of random splits of datasets (Søgaard et al., 2021): to simulate real-world drift and build a better estimation of actual system performance, one should pursue biased sampling of test data, along dimensions such as sentence length, or chronology which is especially impacted by language evolution. This is one more area where NLP standards could come into play, by establishing lists of relevant dimensions to account for when making such choices for an NLP system.

Regarding interpretability and explainability, while some use those terms interchangeably, others have drawn firm distinctions: interpretability is “loosely defined as the science of comprehending what a model did (or might have done)” (Gilpin et al., 2018), while “Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand” (Arrieta et al., 2020), thereby putting the cognitive load of that understanding process more on the model and less on the human. As it seems that “interpretability” has become the preferred term in the NLP community, while other fields rather use “explainability”, should that difference of focus be understood as a conceptual divergence of interests (whereby the NLP community would foster more involvement of the human in model understanding than other communities), or only as a terminological discrepancy? Concurrently, “explainability” as expressed by some other audiences (especially non-technical ones) has nothing to do with either of those concepts, and is rather a synonym for “transparency”, “testing”, or even “reproducibility” (Brennen, 2020), hence a looming crisis if policy makers mean one while practitioners understand the other. As for transparency, while there is consensus on the need for auditability and documentation, the question of what has to be documented and how is still open (Saxon et al., 2021).

7 On the benefits of NLP standardization

Looking at the long-term impact of the Universal Dependencies project hints at how standardizing NLP could more generally benefit the field and its dynamics. The immediate benefit was a more faithful evaluation across languages, that enabled deeper investigation of cross-lingual transfer. But the existence of common guidelines also created a community incentive to produce more data, by providing the guidance and means for easier extension to dozens of zero-resourced languages. The creation of the project itself opened new fora for community-level collaboration and sharing, thereby giving dependency research a new boost. This has been an opportunity to reflect upon research practices, fostering more systematic studies on annotation scheme impact, better highlighting the gaps in linguistic coverage, and uncovering biases in our view of syntax. Overall, standardization contributes to better driving research, both at the individual and institutional levels. A clear taxonomy makes it easier to identify scientific gaps, more compatible resources and tools offer richer experimental means, and shared definitions guarantee that we are all pushing in the same direction.

Standards also support community-wide adoption of good practices. Even if abundantly discussed and well documented as checklists, especially regarding evaluation (Ribeiro et al., 2020; van der Lee et al., 2019; Gehrmann et al., 2022; Marie et al., 2021; Escartín et al., 2021) and documentation (Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019; Ligozat and Lucioni, 2021; Wilkinson et al., 2016), consensual good practices guidelines are not necessarily implemented in practice, in research and even more crucially in industry. Escalating them to formal standards makes it easier to enforce them.

Comparability is another clear benefit for NLP researchers, but even more so for users and consumers. Interoperability can facilitate putting a researcher’s ideas into users’ hands, with easier integration into products. But it is also a matter of survival for SMEs, for packaging and distributing their products in a competitive environment where Big Tech standalone solutions dominate the market and SMEs struggle to propose large-scale alternatives.

Last but not least, standardizing NLP concepts is a necessary step to refine a shared roadmap to address ethical considerations in NLP (Hovy and Spruit, 2016; Leidner and Plachouras, 2017) – but it is also a prerequisite to regulating abuses and enforcing safe use of NLP that preserves individual rights. At a time where the EU is establishing its AI Act, it has started collaborating with CEN-CENELEC to bring clarity on the terminology and processes needed to legally ensure key trust characteristics, such as robustness, trans-
parency or fairness. Standards used to support compliance with the AI Act will thus be written by CEN-CENELEC/JTC 21 over the next few years. In that context, it is crucial that NLP standards (not only AI generic standards that do not fully apply to NLP specificities) are developed, to ensure that companies distributing NLP products in Europe are able to comply with the regulation.

8 Last words: are we ready?

This review of various aspects of the NLP ecosystem has shown how the lack of standards can cause confusion, inefficiency, and sometimes even render research efforts detrimental to scientific progress, through misinterpretations and fostering of bad practices. Building and formalizing consensus on key practices and concepts would enable instead a more reproducible, more insightful, more industry-ready and more ethical science.

Admittedly, not everything can be readily standardized: sometimes the scientific material to do so does not even exist yet. And sometimes research freedom and creativity need to be preserved by maintaining concurrent options. But these are precisely cases where it is even more important that the standardization ecosystem benefits from scientific expertise, in order to avoid over-standardizing the field, or widening the discrepancies between research and industry practices.

We believe it is a matter of scientific responsibility to offer such guidance to those who are shaping the industrial and legal future of society-wide use of NLP. Contributing to standardization means sharing our expertise and insights, but also our needs and our concerns, both as scientists and as citizens. NLP is ready and in need – now we have to get ready.

There are numerous ways to taking part. While community-internal initiatives should be pursued and fostered, we also encourage European researchers to join CEN-CENELEC/JTC 21\(^6\) for contributing to its budding roadmap, and worldwide researchers to both pursue resource standardization efforts within ISO/TC 37\(^7\) and help ISO-IEC/JTC 1/SC 42\(^8\) to deepen debates that have still only scratched the surface of the upcoming work. Organize events, discuss, share, debate, draft, brainstorm, publish. And NLP standards will be within reach.

Limitations

A significant part of this paper has a purely illustrative value, and the provided set of examples does not convey a comprehensive view of the existing standardization issues. Similarly, despite extensive search, we offer no guarantee of exhaustivity in our inventory of NLP standardization groups, in particular for non-cited SDOs (e.g. IEEE).

The review and discussion are also biased towards a number of European concerns and initiatives, which may be either a symptom of its pioneering position on the topic, or merely a lack of depth in our survey of local initiatives in other parts of the world. National-level standardization efforts are not discussed either.

Finally, this work only scratches the surface of discussing the scientific and industrial feasibility of standardization for each part of the field, which may significantly vary from one task or concept to another, depending on their maturity and history.

Acknowledgements

This work has been partly funded by a StandICT.eu 2023 Fellowship. We warmly thank members of JTC 21’s ad-hoc group 4 for fruitful discussions, and in particular Rania Wazir who kindly reviewed a preliminary version of this paper. We also thank Thomas Zielke for inspiring this paper, and Laurent Romary for his long-term involvement in language resource standardization.

References


Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. The Gravett, 14.


ELSE. 1998. Towards a European evaluation infrastructure for NL and speech. Workshop at LREC.


NLP Open Source Software (NLP-OSS), pages 1–6, Melbourne, Australia. Association for Computational Linguistics.


Ralph Grishman and Beth Sundheim. 1996. Design of the MUC-6 evaluation. Technical report, NEW YORK UNIV NY DEPT OF COMPUTER SCIENCE.


High-Level Expert Group on AI. 2019. Ethics guidelines for trustworthy AI.


Mike Kroutikov. 2019. 7776 ways to compute F1 for an NER task.


Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency, pages 220–229.


Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering, 27(2):443–460.


