Pre-print of full paper presented at CONVERSATIONS 2023 – the 7th International Workshop on Chatbot Research and Design, hosted by the University of Oslo, Norway, November 22-23, 2023. The final version of the paper will be published in the forthcoming Springer LNCS post-workshop proceedings.

Language Ideology Bias in Conversational Technology*

Sviatlana Höhn[1 [0000−0003−0646−3738]], Bettina Migge[2 [0000−0002−3305−7113]], Doris Dippold[3 [0000−0001−6193−4710]], Britta Schneider[4 [0000−0003−0377−5808]], and Sjouke Mauw[5 [0000−0002−2818−4433]]

[1] LuxAI, Luxembourg sviatlana.hoehn@luxai.com
[2] University Colledge Dublin, Ireland bettinamigge@ucd.ie
[3] Universtity of Surrey, UK d.dippold@surrey.ac.uk
[4] European University Viadrina, Germany bschneider@europa-uni.de
[5] University of Luxembourg, DCS, Luxembourg sjouke.mauw@uni.lu

Abstract. The beliefs that we have about language are called *language ideologies* and influence how we create and use language technologies. In this paper, we explore language ideologies and their role in the process of language technology design using conversational technology as an illustrative example. We draw on two qualitative studies, both of which aim at discovering common language conceptualisations in the context of language technology design through collaborative work with study participants. In study 1, we use a survey, group discussions and co-design methods with technology developers. In study 2, we use a survey and group work with technology users. We found that standard language ideology is intertwined with a referential (language in its function to convey information) view on language data in the development process, and that a conceptualization of language as referential tool dominates the language technology landscape. However, participants in both qualitative studies are aware of other functions of language. Further we found that language ideologies are intertwined with public discourse about language technology, and upcoming policies on AI regulation will reinforce these ideologies. We argue that non-referential functions of language must be integrated into language models, and that the actual practices of both language and language technologies must be carefully considered for improved conversational AI and effective policies.

Keywords: Language ideologies · Conversational AI · Language functions.

1 Introduction

People have widely differing ideas of what language is, and multiple linguistic theories reflect different views on language and its functions (see for instance [1]).

* Supported by FNR Luxembourg INTER-SLANT 13320890, IF CAIDA 17762538, and EU COST Action LITHME CA19102

Language technology developers typically seek computationally implementable definitions of language [21] and collaborate with linguists in numerous multidisciplinary projects. In this process, they inevitably draw on cultural conceptions of language, which are called language ideologies [10]. In the simplest formulation, language ideologies are culturally conditioned sets of beliefs about language that are not necessarily grounded in linguistic or empirical evidence but intersect with social belonging and social hierarchy [34]. Language ideologies influence how language technology creators design and develop the technology that people use for communication, and how users use and perceive language technologies. They concern, for example, language standards and referential functions of language. The former postulate that language is organised according to national categories (e.g. English, Polish and German), and that national or ethnic groups speak the same language that has one and only one correct form. This is often associated with the view that some linguistic practices have a higher value than others and that language change is negative. Referential ideologies construct language as a code that transmits referential, non-linguistic content, and regards meaning as a priori and inherent in words and sentences. Language is argued to exist on its own, while humans use it on demand [27].

Yet, researchers' input on non-referential functions of language shows that language has many more functions (such as expressing and creating relationships, or negotiating and creating mutual understanding). Capturing these functions in computational approaches to language can help to improve systems based on language technology. For example, taking language variation, style variation and language change into account improves word prediction, text classification systems [22], bias detection [9] and speech recognition [36], and it supports social science research [16]. In addition, it has been argued that inclusion of minoritized or non-standard languages in language technology may add to the prestige of these languages, and guarantee citizens' language rights, reduce social inequality, and narrow the digital divide [18].

Language ideology-related problems are especially visible in conversational technology. Research about user satisfaction has found that users of conversational technology have different expectations of language style, e.g. formal vs. informal, which demonstrates the relevance of non-referential variation of language in chatbot use [29, 19]. Yet, chatbot interactions tend to be designed according to standard language and entail referential ideologies. This has the effect that users of non-standard language are marginalized and that sociolinguistic hierarchies between "good", "normal", "useful" and "deviant", "bad", "impractical" language are reproduced. Furthermore, the fact that referential meaning is an outcome of social interaction and that language therefore constantly changes is rendered invisible, as language tends to be treated as referential, stable code.

Conversational technology has a potential to reinforce negative attitudes in humans towards non-standard language variations, for instance towards African-American Vernacular English [3]. Conversational Artificial Intelligence (CAI) also caused political debates every time a prominent technology producer made technology available that was not yet mature enough, such as chatbot Tay by

Microsoft, [33], or when some features of the conversational technology reached a milestone in the simulation of human behaviour which made it difficult to differentiate machine from human activity, such as Google Duplex [24]. The most recent development that caused intensive academic and political debates and amplified public discourse is ChatGPT by OpenAI [26].

In light of the above, it is important to model social factors in language technology [14]. More specifically, non-referential functions of language such as identity construction, setting group boundaries and regulation of social proximity need to be made part of language models to improve political bias recognition [15]. In addition, including language variation into security mechanisms of multilingual large language models can help prevent prompt injection attacks more effectively [12]. However, language ideologies play a role in both, CAI development and CAI-related planned regulations. Thus, considering the actual use of the CAI systems, real users' needs and language and communication practices is essential to develop safety measures for AI.

This article argues that both the language-ideological as well as the technology-regulatory debates insufficiently consider real users and real-life language use. We investigate the role of language ideology in language technology by looking at conversational technology, although language ideologies influence all types of language technology. In addition, we investigate the relationship between language ideologies and technology ideologies by looking at what problems users face when using language technologies and how they prioritize further improvements. Our **research questions** are:

1. To what extent do conversational technology designers, developers and users express standard language and referential language ideologies?
2. What other kinds of (language) ideologies are identifiable in the conversation technology domain?

To answer the research questions, we conducted two qualitative studies: a two-hour in-person workshop collocated with a European academic venue focused on chatbots (Study I), and a three-hour in-person workshop collocated with a Summer school funded by a European research agency (Study II). Two researchers conducted the workshops: one working in conversational technology and one working in sociolinguistics.

Both studies show that language ideologies affect the entire process of language technology development, from data collection, design, specification, development and testing phases, to user-based evaluation studies. The second study also discloses that political discourse about artificial intelligence in general and conversational AI in particular shift the focus of language technology workers and users from the actual users' needs to politically amplified topics.

This article aims to raise awareness among their designers, developers and users about the nature, effects and role of language ideologies in the life cycle of language technology. We make a step towards opening a public discussion about language ideologies, their incarnation in technology and policies, and the feedback loop to language of those.

2 Language Ideologies, Policies and Public Discourse

All languages are, in terms of grammatical complexity, equal but socially differentiated. Languages associated with powerful nations and print media have higher prestige. Such social hierarchies are observable locally at the level of national language standards, and globally as, for example, in the dominance of English. Language technology reinforces these hierarchies. For instance, speech technology performs better with prestigious Englishes than for less widely spoken ones [20, 5].

Standard languages are perceived as neutral codes, also referred to as "voices from nowhere" [10, p.7]. However, they result from socio-historical circumstances, such as the development of nations, print culture and social hierarchies [4, 13, 35]. Non-standard languages (e.g. regional or social dialects) have a different history, have covert prestige and are given different functions [17]. The third wave of sociolinguistics and qualitative approaches to language has foregrounded language as social practice and highlights its co-creational social function (social boundaries, identities, stances and social communities), stressing the importance of the non-referential functions of language [8].

Existing computationally-oriented venues dealing with language variation[6] contribute remarkably to computational modelling of language variation. However, these rely on distinguishing standard language from "dialects" and the view that language is an external, static artefact. Overall, language technologies could potentially play an important role in destigmatising minoritized languages and the identities associated with them. However, to date language technologies are mostly developed for major languages that have undergone standardization and are associated with a state and a nation. Other languages and varieties are declared as "non-standard", "noisy" and "non-canonical". [7]. Yet, the "messy" nature of language is central for it to (re)create innovative social meanings; it is not a design fault. With the growing importance of conversational technology, it is vital to counter an understanding of language as a referential, denotional resource [31] and recognise interaction as social practice that negotiates the non-referential functions of language, such as identity construction, face work, group belonging and boundary negotiation [30, 32, 28].

Language ideologies are closely intertwined with language policies as they shape and inform the way societies perceive, value, and regulate languages [25]. They influence the formulation and implementation of language policies. Language policies, in turn, are the concrete actions and rules put in place by governments, institutions, or communities to manage language use and distribution. The alignment or conflict between (some) language ideologies and language policies can have profound consequences, impacting linguistic diversity, social inclusion, and access to resources. Understanding this complex relationship is crucial for promoting linguistic equity and fostering inclusive societies.

Policies related to regulation of language technologies can work in the same way as language policies. Policies related to language technologies will have a sig-

[6] E.g. https://sites.google.com/view/vardial-2022/home

nificant impact on both the development and usage of language technologies, as well as on the languages themselves. These policies encompass regulations, standards, and guidelines that govern the creation, deployment, and accessibility of language technologies. They also determine the nature of language technologies and how they are integrated into various sectors, affecting industries, education, and government services.

Current summaries of the planned big set of policies related to AI in general and language technology in particular[7] suggest that planned policies are very much influenced by public discourse which, in turn, is shaped by new ideologies and beliefs about language technologies, and not necessarily by their actual use. Effective public discourse can positively influence policy-making, shape policy outcomes, and hold policymakers accountable, making it a fundamental component of a democratic and participatory society. However, public discourse about language technologies based just on ideologies (fed by the standard language ideology which postulates the existence of a neutral code) in the absence of attention to the actual practices of users (which crucially involve language choices aimed at embodying social meanings), will misinform policy makers and lead to sub-optimal, ineffective and unpopular policies.

This article seeks to raise awareness of these issues and proposes steps towards changes.

3 Method

As mentioned in Section 1, we conducted two qualitative studies including two surveys and two workshops aiming at disclosing language ideologies in language technologies. We conceptualise the participants of Study I as language technology producers (designers, programmers, researchers), and the participants of Study II as language technology users. Both groups represent highly skilled and highly educated European populations. The first workshop included three steps:

1. *Pre-workshop data collection.* The participants of the workshop were invited to fill in an optional survey with 30 questions. The topics covered the participants' concepts of quality data and language variation, language and language use, interaction and its purpose, and concepts of linguistic disciplines. The information was elicited via open questions. Twelve responses were analysed qualitatively for Step 2.
2. *In-workshop discussion.* Participants introduced themselves. Facilitators gave an introduction about language ideologies and discussed the survey results. The results were presented by a computer-science researcher and commented by a sociolinguistic researcher. This phase helped to answer the first research question.
3. *In-workshop co-design.* Participants discussed three questions in smaller randomly assigned groups and presented the results in the plenary session before jointly formulating mitigation and awareness-making measures.

The second workshop was composed of four stages:

1. *Pre-workshop data collection.* The participants were invited to answer ten questions about their views on language and language technologies.
2. *Exploratory use case.* The facilitator presented a conversational robot and explained its use case: teaching children different types of skills including language. The participants were invited to write down on sticky notes their ideas about positive and negative effects with respect to one of the categories: language rights, language ideologies, language variation, interaction, language diversity and vitality, language learning, language technology and language work. The participants were also asked to place their sticky notes on a whiteboard according to the category for which they think the statement is relevant.
3. *In-workshop group work and discussion.* Facilitators presented the results of the survey. Based on the preliminary outcome of the survey, language learning apps were the most known type of language technology. Therefore, the participants were asked to design their own language learning application that solves all problems detected by the survey.
4. *Cross-domain analysis.* The researchers analysed and aligned all outcomes produced by participants: survey, group work and ideas from sticky notes.

Both groups were mixed in terms of gender and academic seniority level and included expertise in linguistics, language teaching, language translation, interaction, dialogue analysis, user experience research, psychology, behavioral sciences, computer science, language technology, communication and media studies. The number of subjects is 22 for the first and 43 for the second group.

The results of the qualitative questions from the surveys were independently coded by both researchers who conducted the studies. If multiple categories were identifiable in the answer, the answer received multiple labels. Examples of coding are provided in Table 1.

Question formulation	Answer example	Labels
In your view, which problems of language technologies should be urgently resolved and why?	I am not sure if all problems with language technologies can be resolved. What is most urgent is for there to be a more honest and comprehensive narrative that teaches the general public about the benefits of language technologies and their drawbacks...	Literacy
Same as above	data security issues; educate laypeople about problems language techologies can bring	Privacy Literacy

Table 1. Examples of coding for qualitative questions in both surveys.

4 Results and Discussion

In this section, we describe our findings from both studies in detail and synthesize the results with regard to language technology related ideologies.

4.1 Study I

This qualitative study confirmed the language ideologies and the participants' focus on referential functions of language. The results allow drawing a number of conclusions about state-of-the-art in the language technology domain.

The study also revealed that there is a connection between standard language ideology and a referential view of language. Standardised, normalised language data make the mapping from input to meaning representation easier, but it is not always made transparent, how data "normalization" affects this process (see for instance [23] for a discussion). However, the need for "clean" and error-free data is problematic because even standard language is not uniform across social domains, people use and process language idiosyncratically and creatively, and constantly adapt language in interaction for self presentation, to encode their views about the interaction, the interlocutor etc. Thus, non-referential meanings are erased in data normalisation and cleaning.

Linguistics and sociolinguistics have made considerable progress since the emergence of generativism, however, their findings have rarely been included in the design paradigm of language technology [2, 11]. Moreover, co-creation of language *while* interacting with language technology has never been conceptualised due to standard language ideology dominating in the language technology community which maintains that language is neutral and referential in nature. Technical constraints are central in research on language technologies. The details are presented below.

Findings from the Survey I The survey helped to disclose A) the participants' assumptions about language, its function and variation, B) their beliefs about communication, and C) their perspective on language data.

A) The responses disclose participants' awareness of multiple forms and functions in each language, and the way these encode different aspects of context. The participants associated every-day language with emotions, a playful, familiar, friendly manner, and social closeness with accessible vocabulary and low complexity of expression. Standard language, in contrast, is associated with respectfulness and social distance. Overall, the more private the space, the more diverse are people's language choices. Participants mostly use English for communication in their professional contexts even if it is not the language for which they develop technology. Other languages dominate private communication. Participants were knowledgeable about types of language variation (interactional variation, dialectical patterns, syntactic and lexical variation) although it is usually not part of their language technology design or development process.

B) Communication was named as the main function of language and described as the transmission of thoughts, feelings and experiences. We identified three clusters of responses related to the nature of human communication: a *functional* cluster that included sense of purpose, ability to do repair, create relevance and manage the unexpected; an *emotional* cluster that included expression of emotions and empathy, humour and sarcasm; and a *context* cluster that included ambiguity, context dependency of meaning and spontaneous and

endlessly variant production. All these attributes characterise different aspects of the referential function of language. An awareness of the non-referential functions of language had low salience in the responses (except of the emotional cluster). The latter was mostly associated with explicit labelling of emotion using positive and negative words, emphasising the referential function. There was lack of awareness that language emerges *in* the interaction, and that language is always intertwined with the social, and cannot be separated from it.

C) Although there was a high demand in the group for naturalistic data, such as human face-to-face conversations, the researchers typically choose "clean", well documented, standardised, error-free and accessible data for their work in language technology. Thus, the *operational* (prioritising ease of implementation) view on data for language technology dominates the community. In this way, the standard language ideology bias is amplified in language technology: language data are taken from the standard language and modified to be error-free, "cleaned", well-documented etc.

Because the desired data are usually not available in the needed quality or quantity, participants reported that they usually create their own datasets via surveys, log analysis, experimental designs or use "similar" data (e.g. Twitter comments instead of instant messenger chat logs).

Findings from the group work I We asked every subgroup to draw their position on the following three questions:

Q1 Which social functions of communication does chatbot design try to model or implement?
Q2 Which understanding of language underlies chatbot design?
Q3 What is the relationship between data, data collection and generation methods and chatbot's communication capabilities?

Each subgroup produced a poster, and the results are summarised in Table 2 by question (left column). Although the group was aware of the relationship between language data and language technology based on them, the group work again highlighted the operational view on language data for language technology development, see Table 2 Q2 ("standardize it", "input - output"). Language data are seen as a help to find the right mapping from input to output, equally for standard and every-day language, for task-oriented and social dialogue. The referential role of language is not only very dominant, but also very convenient for implementation, which makes it difficult to change the paradigm.

Findings from the co-design phase In the co-design phase we aimed at formulating language ideology awareness guidelines together with the participants of the workshop. The following points were emphasised by the participants.

Co-design with users. Involving users into the process of language technology design from the beginning can help to go beyond solely referential and operational views on language in conversational technology and language technology in general.

	Group 1	Group 2	Group 3
Q1	Formal language (as it has rules + less creative). Bots seem more human when using puns and jokes.	Activate, engage; interpersonal understanding, empathy; exchange information; showing manners, etiquette, scrutiny smalltalk; humour; building intimacy, trust, "talking"; communicative entity is social to begin with.	Standard language - transactional, goal-oriented, closed-domain; everyday language - relationship building, trust, empathy, compassion, personification, adaptation; open-domain.
Q2	Standardize it, so it can learn from a big data sample, then it does not adapt to each individual (language-wise); Oral communication → pitch, tone can play a role in language meaning, text-based needs different cues to implement this. It is not the understanding but the technical ability to write concisely (short sentences, to the point).	Aim to objectively understand language (encode, decode), context-specific (domain), modelling (transforming voice to text, from 3D to 2D), adaptation (reduce language complexity).	Operational understanding of language, defined by Q1.
Q3	Garbage in, garbage out; drawing of mutual dependencies of relationships.	Drawing: collection methods influence quality, data quality determines capabilities. What defines quality? More data is not always better. Biases. Questioned direct relationship between collection methods and a chatbot's capabilities.	Data ethics, privacy, data governance, data ownership, transparency, context-dependent.

Table 2. Transcript of the group work, original wording.

Detailed documentation. Transparency about the gap between "natural" interaction data and the actual data that are used. The origin of data must be documented. It must be made explicit to the readers of documentation for datasets, what kinds of approximations were made for each application case.

Continuous improvement. Starting with a simplified "minimum viable product", design and development of language technology is a step-by-step, iterative process. Simplifications need to be made explicit to the readers of documentation, and conversation design is never complete.

4.2 Study II

The study revealed that there is a relationship between language ideologies and political discourse about language technologies and their regulations. The survey

disclosed that language technology users are clear about their actual needs and have a very concrete understanding about how language technology would make their work and daily life more efficient, pleasant and enjoyable. However, when asked about priorities of language technology improvement, actual user needs are downgraded and publically much discussed topics topics are prioritised.

In addition, this study revealed that participants perceptions of specific language technology is also determined by public discourses. Both the use case and the group work echoed the public discourse.

Findings from the survey II The attendees of an in-person Summer school were asked to fill-in an optional survey. From 43 attendees, 19 chose to reply to the questionnaire. The findings were discussed with the entire group.

The survey compared for which purpose the participants use language technology, what participants viewed as the biggest problems related to their own technology use and experience, and what should be improved as soon as possible. The top five in these three categories were:

Problems	Benefits	Priorities
Low quality	Productivity	Literacy
Language availability	Creativity	Inclusion
Security & privacy	Progress	Bias
Expectation-quality mismatch	Facilitate communication	Language coverage
Confidence-quality mismatch	Accessibility of information	Ethics

Further, we compared familiarity, frequency of use and interest in language technologies by participants. Top five for these three categories were:

Familiarity	Usage	Interest
Machine translation	Messengers	Translation
Speech synthesis	Machine translation	Language learning, teaching
Language learning apps	Authoring tools	Research
Speech recognition	Social media	Everyday life
Spelling & grammar checker	Language generation	Communication

As in Study I, the participants acknowledge that language performs a variety of functions in human-human communications. However, the functions of language in communication with technology are in 80% of the cases reduced to commands and instructions (referential view).

Besides communication technologies such as messengers and social media, machine translation and text production tools (text authoring and language generation) are the most popular technologies, and their relation to the top five benefits listed by our participants is quite clear. It is also quite clear that low quality and limited language coverage (language availability) of such services leads to frustration with technology. However, literacy related to language technologies, inclusion, bias and ethics that are listed among the top priorities for future development will not help to solve problems related to low quality.

This mismatch in responses suggests that even technology users set technology development priorities based on public discourse rather than their own needs.

Findings from the use case The facilitator demonstrated a social robot that is used for teaching autistic children different types of skills. The demonstration focused on language technologies used to deliver such kind of educational interventions, and mutual dependencies among them (dialogue design, automated speech synthesis, feedback generation and so on).

The task was to pin down all positive and negative aspects of the demonstrated technology by writing statements on sticky notes. Those sticky notes were first categorised by participants according to the eight categories provided in the beginning (see Section 3), and later coded and clustered by the facilitators according to their subject. Participants listed 75 negative and 55 positive aspects in total that were coded to 25 and 13 categories, respectively.

With regards to language ideologies, this part of the study showed that participants' awareness related to language ideologies is rudimentary. In total, eight statements were categorised by participants as related to language ideologies, and only three of them were in fact related to language ideologies, all addressing the robot's potential to promote and reinforce some standard language. At the same time, language variation and "human-like" interaction are seen as the most urgent problems with social robots, although these problems have been classified by participants as related to language variation.

The top five of issues listed are:

- language coverage and variation (15);
- ethics (12) including replacement of humans by machines (10) although in the described scenario, no replacement took place;
- technical accessibility and price (7);
- issues with interaction (7); and
- potential technical problems (5).

Top three positive aspects listed are related to human empowerment, interaction quality and enrichment of teaching and learning followed by four categories with the same strength: accessibility, language coverage, personalisation of interaction and potential other application scenarios.

The comparison of positive and negative aspects shows that there are no absolute values: the same technology is seen as empowering and disempowering humans, restricting and providing opportunities for a wide language coverage and accessibility, offering positive and negative interaction experience. This shows that how people judge technologies is very much dependent on how messages are framed, also in public discourse.

Findings from the group work II The following task was given to participants after the facilitators had presented the results from the survey and given a theoretical introduction on language ideologies.

Imagine, you need to design a new language learning app and to do so, you need to solve three questions first:

1. *Which social functions of communication should the new app implement?*
2. *Which understanding of language should underlie the language learning app design?*
3. *What kinds of data should this design process rely on and how can this data be obtained to meet the targeted social functions?*

Participants had 15 minutes to work in small groups on these questions and were then asked to present their solutions. The facilitators were interested in how clear those tasks are to the groups and how successful the groups are in their solutions. Participants worked in six groups and each group included experts in different domains and different levels of academic seniority. All six groups struggled with understanding and definition of social functions of language and language as a whole when it comes to design of a concrete technology. None of the groups could clarify the question about data.

Findings from the cross-domain analysis The main three outcomes from the survey, case study and group work of Study II are:

1. When it comes to an abstract discussion of language technologies, there is a mismatch between language technology user needs and improvement priorities set by the same users. This mismatch is likely due to public discourse related to language technology.
2. When it comes to a discussion of a specific conversational technology, positive aspects are usually those mentioned explicitly in the demonstration (empowerment, interaction, enrichment) while negative aspects are related to current public discourses (bias, ethics).
3. When it comes to designing solutions, multidisciplinary teams of experts still struggle with finding a proper working definition for basic concepts such as language and language functions. There is also little clarity about the types of language data that exist, their suitability for technology design and how they may be acquired.

4.3 Discussion and Potential Actions

This article presents qualitative research obtained from two small-scale qualitative studies. The initial ideas generated through the questionnaires and the workshops need to be validated further through an interview-based study with developers, investors and policy-makers, or a survey distributed to a larger sample of respondents.

Study I indicates that language ideologies play a role in the entire life cycle of language technologies, while Study II suggests that larger social ideologies such as current public discourses on technology and beyond, as well as their drawbacks, also affect how we imagine, create, use and evaluate language technologies. Both language ideologies and public discourses about language technology heavily impact our views of technology and its regulations and contribute to sidelining the consideration of users' needs and real-life practices.

As mentioned in Section 1, multiple CAI tools ignited controversial public discussions about ethics in CAI systems. ChatGPT in particular raised a lot of questions related to text authoring, teaching, manipulative content production, colonial methods in the process of data acquisition and cleaning, and intellectual property rights in general [26]. At the same time, the European Union started preparing the EU AI Act, a document that is supposed to provide guidelines for EU member states on how to handle particular matters in the national legislation. One of the main claims that the EU AI Act draft is making is that all texts generated by generative language models must be marked as generated.[7]

Discourses about language and language technologies are dominated by ideological claims. Policies connect both in the way that language policies have little to do with the actual language practices [25], and the ongoing public and political debates about regulation of language technologies (and AI in general) have little to do with the human practices of using them.

For example, the summary of the planned regulations contains the claim that in the future, all texts generated by language models will have to be marked as such. This requirement will cause multiple practical issues related to human and human-machine collaborative text production practices. Generative language models are used nowadays in spellcheckers, style-checkers, grammar correction features of most text editing programs, in word suggestions on mobile devices, response suggestions in text messengers, and in sentence completion suggestions of text authoring programs. There is a lot of text generation outside of ChatGPT. Further, even if language models such as ChatGPT are used, texts are usually generated in collaboration with human co-authors (reformulations, prompt engineering) and not completely autonomously. It is not clear how much of the text must be generated in order to be declared as automatically generated, and how the amount can be measured in practice.

Both types of ideologies, the ones related to language and the ones related to language technology, will set constraints and judge together, what is a "good" language and what is a "bad" or "low quality" language. However, in addition to labels associated with nation state and power, such as "standard English", labels associated with participation of technical devices in language practices, such as "automatically generated" will be used to make differences between more valuable and less valuable language. Finally, by regulating language technology without looking at their actual use by humans and without capturing new practices of language creation by human-machine teams, policy makers will set regulations to language itself, reinforcing existing language ideologies in their incarnation in current language technology.

To make the results of this research applicable in practice, we suggest the following actions:

1. **Theory**. Sociolinguistic and pragmatic expertise and more recent theories of language that see it as a social practice must be adopted in language modeling to handle non-referential functions of language. In addition, they

[7] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

must also inform policy-making about the role of technology as not-just-tool in the process of language creation.
2. **Policy**. Work with politics and NGOs to declare availability of language technology in regional variations a human right.
3. **Capital**. Incubators could support new businesses targeting smaller languages. We cannot change capitalism with this paper, but collaborative efforts are needed to emphasize economic and cultural value of smaller languages.
4. **Investment**. Push public funds into language technology for people: focused on user needs, created with potential users, speaking the user's language as messy as it is, and being able to deal with variations.
5. **Data**. Create and use real spoken language resources (no, movie dialogues are not spoken language, but [6] compiled a list). Create more of those corpora.

5 Conclusions

Both studies have several implications for the language technology community and policy makers. First, specialists involved in language technology creation need to sharpen their awareness of the kinds of biases that impact the whole life cycle of language technology. Both standard language ideologies and political discourses on the benefits and drawbacks of technology influence the creation, conceptualization, production, evaluation and further development of technologies (including users' expressed needs) and thus need to be critically assessed and reflected upon at each step in the life cycle. Second, discourses and views of participants in the life cycle of technologies are contradictory and varied because not all participants engage in the same discourses and ideologies. Third, while the standard language ideology continues to be powerful, the language practices it promotes are not always used and in some cases understood only by a minority of people. Fourth, modelling non-referential functions of language is crucial for conversational technology as this is what makes technology appear "human-like". Finally, the public discourse about language and language technology needs to be more inclusive. Instead of echoing claims related to ethics and bias, the community needs to have a wider public discussion about the actual uses of language, language practices, variations, and their relation to technology and power.

References

1. Auer, P.: Sprachliche Interaktion: Eine Einführung anhand von 22 Klassikern. Niemeyer, Tübingen (1999)
2. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. In: Proc. 58th ACL Meeting. pp. 5454–5476. ACL (2020)

3. Cassell, J.: Socially interactive agents as peers. In: The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application, pp. 331–366 (2022)
4. Coupland, N.: Sociolinguistics: Theoretical debates. Cambridge University Press (2016)
5. DiChristofano, A., Shuster, H., Chandra, S., Patwari, N.: Performance disparities between accents in automatic speech recognition. arXiv:2208.01157 (2022)
6. Dingemanse, M., Liesenfeld, A.: From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5614–5633 (2022)
7. Dipper, S., Neubarth, F., Zinsmeister, H. (eds.): Proceedingns of KONVENS 2016: Processing non-standard data — commonalities and differences. Bochumer Linguistische Arbeitsberichte (2016)
8. Eckert, P.: Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. Annual review of Anthropology **41**(1), 87–100 (2012)
9. Ferrer, X., van Nuenen, T., Such, J.M., Criado, N.: Discovering and categorising language biases in reddit. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 15, pp. 140–151 (2021)
10. Gal, S., Woolard, K.A.: Constructing languages and publics: Authority and representation. Pragmatics **5**(2), 129–138 (1995)
11. Hamborg, F.: Towards Automated Frame Analysis: Natural Language Processing Techniques to Reveal Media Bias in News Articles. Ph.D. thesis, University of Konstanz, Konstanz (2022)
12. Harang, R.: Securing LLM systems against prompt injection. https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/ (2023)
13. Holmes, J., Wilson, N.: An introduction to sociolinguistics. Routledge (2017)
14. Hovy, D., Yang, D.: The importance of modeling social factors of language: Theory and practice. In: The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2021)
15. Höhn, S., Asher, N., Mauw, S.: Examining linguistic biases in telegram with a game theoretic analysis. In: Proceedings of the 3rd MISDOOM. pp. 16–32 (2021)
16. Kellert, O., Matlis, N.H.: Social context and user profiles of linguistic variation on a micro scale. In: Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 14–19. ACL, Gyeongju, Republic of Korea (Oct 2022)
17. Labov, W.: The social stratification of (r) in new york city department stores. In: Dialect and language variation, pp. 304–329. Elsevier (1986)
18. Láncos, P.L.: The role of language technologies in promoting the participation of linguistic minorities in social, political and economic life. Foreign Policy Review (2), 73–87 (2021)
19. Liebrecht, C., Sander, L., Hooijdonk, C.: Too informal? how a chatbot's communication style affects brand attitude and quality of interaction. In: International Workshop on Chatbot Research and Design. pp. 16–31 (2020)
20. Markl, N.: Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 521–534 (2022)
21. Mitkov, R. (ed.): The Oxford Handbook of Computational Linguistics. Oxford University Press (2003)

22. Nguyen, D., Doğruöz, A.S., Rosé, C.P., De Jong, F.: Computational sociolinguistics: A survey. Computational linguistics **42**(3), 537–593 (2016)
23. Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., Winters, J.: How we do things with words: Analyzing text as social and cultural data. Frontiers in Artificial Intelligence **3**, 62 (2020)
24. O'Leary, D.E.: Google's duplex: Pretending to be human. Intelligent Systems in Accounting, Finance and Management **26**(1), 46–53 (2019)
25. Pennycook, A.: Language policies, language ideologies and local language practices. The politics of English: South Asia, Southeast Asia and the Asia Pacific pp. 1–18 (2013)
26. Ray, P.P.: Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems (2023)
27. Schlangen, D.: Norm participation grounds language. arXiv preprint arXiv:2206.02885 (2022)
28. Shrikant, N.: "it's like,'I've never met a lesbian before!'": Personal narratives and the construction of diverse female identities in a lesbian counterpublic. Pragmatics **24**(4), 799–818 (2014)
29. Skjuve, M., Følstad, A., Fostervold, K.I., Brandtzaeg, P.B.: A longitudinal study of human–chatbot relationships. Int. J. of Human-Computer Studies p. 102903 (2022)
30. Sowińska, A., Dubrovskaya, T.: Discursive construction and transformation of 'us' and 'them' categories in the newspaper coverage on the US anti-ballistic missile system: Polish versus Russian view. Discourse & Comm. **6**(4), 449–468 (2012)
31. Trosdal, M.B.: Meaning: The referential function of language. Philippine quarterly of culture and society **23**(3/4), 361–368 (1995)
32. Whitehead, K., Lerner, G.: When are persons 'white'?: On some practical asymmetries of racial reference in talk-in-interaction. Discourse & Society **20**, 613–641 (09 2009)
33. Wolf, M.J., Miller, K., Grodzinsky, F.S.: Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications. Acm Sigcas Computers and Society **47**(3), 54–64 (2017)
34. Woolard, K.A.: Introduction: Language ideology as a field of inquiry. Language ideologies: Practice and theory **3**(11), 1–50 (1998)
35. Woolard, K.A.: Language, identity, and politics in catalonia. Brown J. World Aff. **25**, 21 (2018)
36. Wu, Y., Suchanek, F., Vasilescu, I., Lamel, L., Adda-Decker, M.: Using a knowledge base to automatically annotate speech corpora and to identify sociolinguistic variation. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1054–1060. ELRA, Marseille, France (2022)