Low-resource language

Paolo Valentinelli (Università di Trento/Technische Universität Dresden)*

paolo.valentinelli(at)unitn.it

Abstract

Traditional definitions of language resourcedness fail to capture the structural inequalities embedded in LLMs' pre-training data. While historical classifications rely heavily on sociopolitical vitality, artifacts, and human resources, a language's real-world robustness no longer guarantees equitable algorithmic representation. To address this disparity, this paper proposes a novel tripartite framework that combines four established exogenous criteria with a relative endogenous quantitative metric: a language must hold a *dominant* or *co-dominant* statistical share of an LLM's pre-training dataset to be classified as *high-resource*. Under this new paradigm, languages with strong societal infrastructure but marginal AI data representation, such as Italian and German in LLMs like GPT-3, are accurately reclassified as *few-resource*, while those lacking both real-world and digital infrastructure remain *low-resource*. Despite industry trends toward data opacity, adopting this precise terminology equips researchers with the tools to expose engineered linguistic disparities, challenge the false equivalence of traditional labels, and advocate for genuinely equitable, multilingual AI systems.

Keywords

artificial intelligence, large language models, language resourcedness, training data bias, low-resource language

1 Introduction

For decades, the term *low-resource*[1] has served as a crucial, if imperfect, shorthand in corpus and computational linguistics. It has been used to categorise languages that lack the requisite data, tools, or investment to benefit from advances in natural language processing (NLP). Historically, the definition of a *low-resource language* (LRL) has been context-dependent, evolving from a descriptor of computational scarcity in the 1940s and 1960s to a proxy for data scarcity in the statistical NLP era of the 1990s and 2000s. However, the advent of large language models (LLMs) has precipitated the term into a definitional crisis.

The very concept of what it means for a language to be resourced is being challenged by the unique constraints of the LLMs' pre-training paradigm. Supporting a wide variety of languages is crucial for applications like search engines, voice assistants, and chatbots, as a broad language support allows more

* Special thanks are due to Prof. Dr. Anna-Maria De Cesare and two anonymous referees for their invaluable suggestions, particularly regarding the refinement of the criteria used to classify *high-*, *few-*, and *low-resource languages*. Their feedback not only enriched the paper with a greater plurality of perspectives but also significantly enhanced the efficacy of the discussion.

[1] In the literature, many alternative terms are found, such as *digitally-disadvantaged*, *less-resourced*, *resource-poor*, *under-resourced*, *underserved*, *understudied*, *zero-resource*, among others.

AILing

Valentinelli, Paolo. 2026. Low-resource language.
AI-Linguistica, Vol. 6 No.1
DOI: 10.62408/ai-ling.v6i1.77

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses ISSN: 2943-0070

users to access these tools in their native language (Pakray et al. 2025). Nonetheless, while there are over 7,000 languages in the world (Eberhard et al. 2025), many applications only provide support for a highly restricted number of languages.

Despite LLMs showing impressive multilingual abilities, their pre-training data has been predominantly English-centric. Thus, this paper argues that the current LLM pre-training paradigm has created a new set of hidden constraints that render traditional definitions of LRL obsolete. Furthermore, we posit that even languages with large speaker populations and significant digital footprints, such as Italian and German, function similarly to LRL within the LLM ecosystem. On the one hand, these languages may suffer from systemic underperformance due to pre-training data contamination, architectural biases, and the pervasive influence of English interference. On the other hand, these new constraints are not merely additive for endangered and minority languages. Instead, they compound traditional challenges and lead to a decline in performance and utility.[2] We will take Cimbrian, an endangered Upper German variety in the Trentino-South Tyrol region in Italy, into consideration (cf. Bidese 2021).[3]

This paper is structured in four sections. After the Introduction, Section 2 provides a historical overview of the evolution of resource language definitions, tracing their path from computational constraints to the multidimensional frameworks of today. Section 3 analyses the LLM training paradigm as a new constraint vector, presenting evidence of systemic bias and performance degradation, offering case studies of this phenomenon, first examining Italian and German as examples of "hidden" resource constraints and then exploring the exponential degradation faced by endangered languages, such as Cimbrian. Finally, Section 4 proposes a new, multi-dimensional resource framework tailored to the LLM era, incorporating AI-specific criteria.

2 The historical evolution of language resourcedness

In the initial decades of natural language processing (NLP), spanning from the late 1940s to the late 1960s, the primary constraints on applications were computational, rather than linguistic (Jones 1994: 4). At the time, machine access was extremely limited, processing speeds were slow, and storage capacity was minimal, particularly when judged by modern standards, which created significant hurdles for researchers. Crucially, the problem was therefore understood to be rooted in the

[2] As AI-generated content increasingly floods the web, it dilutes the already sparse pool of human-written data for these languages, creating a negative feedback loop that leads to a rapid decline in downstream performance.

[3] To illustrate this phenomenon, the following analysis examines qualitative examples from Italian, German, and Cimbrian. These languages were purposefully selected due to the author's native and/or academic linguistic proficiencies, which enables a qualitative assessment of the nuanced cross-linguistic interference and syntactic deviations present in the LLMs' multilingual outputs. While these examples serve an illustrative purpose for this conceptual framework, they effectively highlight the varying degrees of algorithmic marginalisation across different *resourcedness* tiers.

available technology and formalisms, not in the languages themselves. As Jones (1994) highlights, the field was centred on the ambitious goal of machine translation, mainly involving Russian and English, with some minor efforts on Chinese, and languages were not yet characterised by the term *low-resource*. This focus on technology continued throughout the 1960s and 1970s, even as the domain broadened to incorporate semantic representation and knowledge bases (Jones 1994: 6). Compared to the sophisticated analysis involved in contemporary machine translation (MT), the input to early systems was limited, and the language processing was very basic.

The 1990s heralded a significant departure from earlier rule-based systems toward statistical methods, a period that Jones (1994: 10–11) described as the "massive data-bashing period". This change fuelled a heightened interest in using corpora to identify patterns of linguistic occurrence and co-occurrence, which in turn could be applied to compute syntactic and semantic preferences. Crucially, this shift had a profound impact on the concept of LRL. As the new models became dependent on large datasets, the vast majority of the world's languages were suddenly redefined as under-resourced, since statistical models could not be effectively trained without extensive annotated text corpora.[4]

Nowadays, pre-trained LLMs have become a cornerstone of modern NLP pipelines because they often produce better performance from smaller quantities of labelled data (Devlin et al. 2019; Radford et al. 2018, 2019). The critical concept is, therefore, pre-training: an unsupervised or semi-supervised learning process where neural networks acquire language understanding. This is achieved by performing tasks like sequence labelling or text classification on vast amounts of textual data (Radford et al. 2018). The usefulness of pre-trained LLMs was subsequently confirmed by the discovery that they could perform valuable tasks without further training. This led to a continuous trend toward increasing scale, driven by the empirical observation that LLM performance generally improves with greater size (Le Scao et al. 2023: 3).

However, the emergence of this data-driven environment creates the fundamental challenge of defining LRLs within the context of LLMs.

> Less than 5 percent of the roughly 7,000 languages spoken around the world today have meaningful online representation, leading to what has been called a "digital language death"—when the lack of technological support for a particular language precipitates its decline. Today, languages that are not fully digitally supported are less likely to be recorded—never mind included in AI model training datasets. As a result, language resourcedness varies widely. (Pava et al. 2025: 8)

[4] It is crucial here to distinguish between *raw data* (plain, unannotated digital text) and *annotated data* (structured text enriched with linguistic, semantic or metadata tags). While raw data for many minority languages may exist on the web, the performance of LLMs depends heavily on annotated corpora and high-quality datasets. These require significant human labour and computational infrastructure, which minority language communities usually lack.

Foundational work, such as Joshi et al. (2020) and Simons et al. (2022), attempted to systematise resource classification by introducing taxonomies based strictly on the availability of labelled and unlabelled data. Figures 1 and 2 illustrate the schematic representations adapted from these studies.

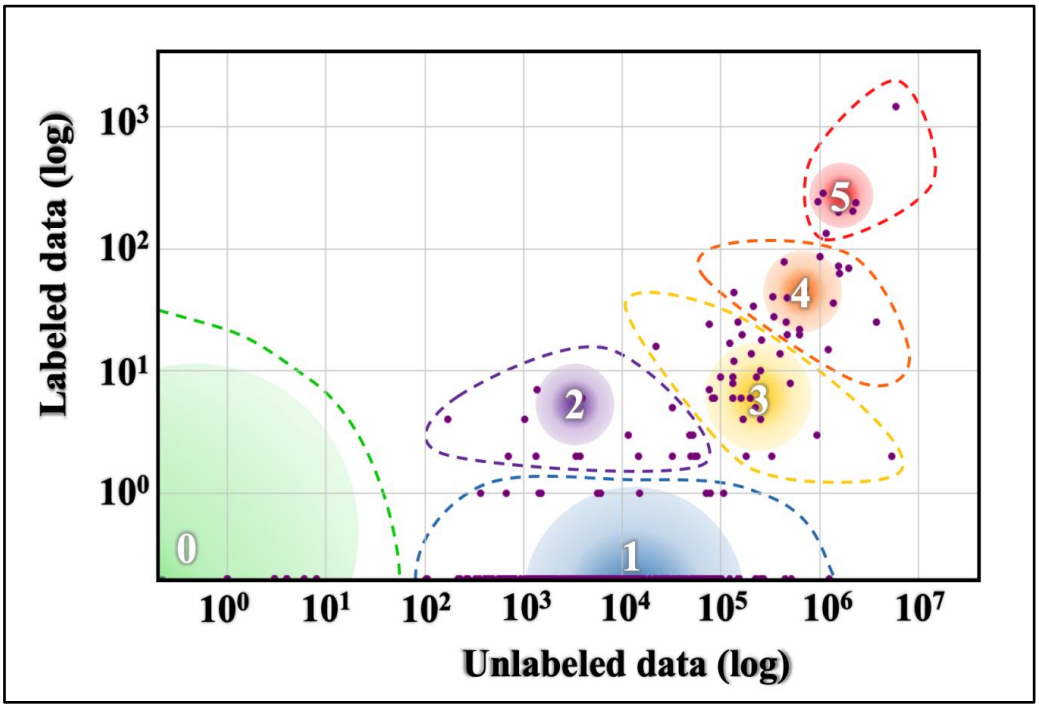

Figure 1: Taxonomy of language resourcedness by Joshi et al. (2020).

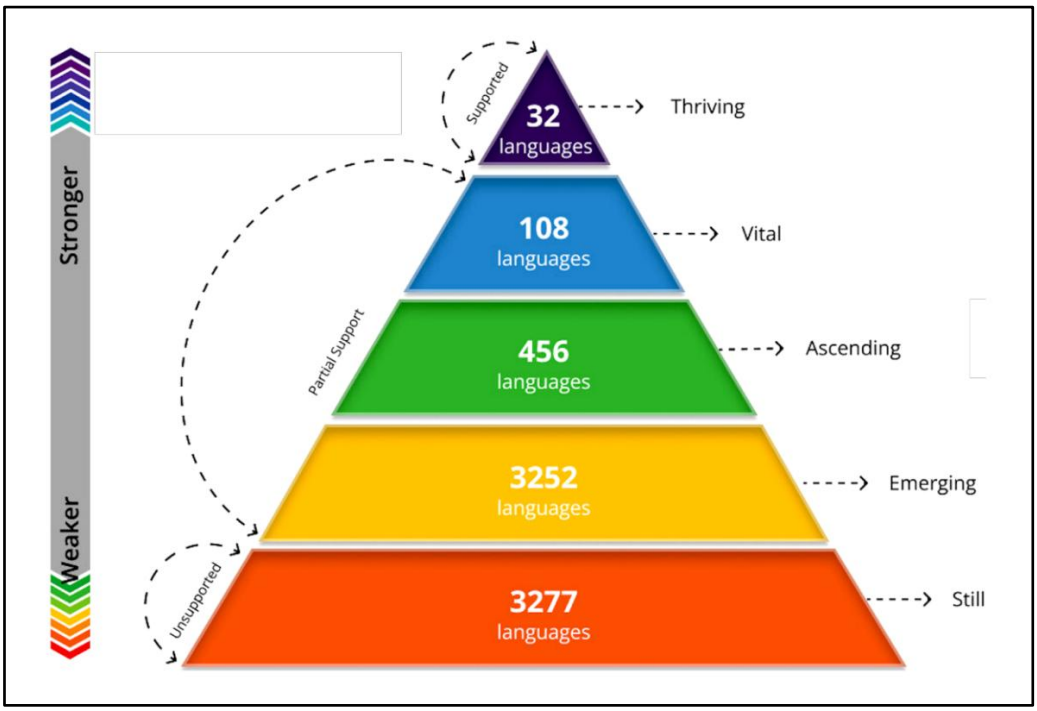

Figure 2: Taxonomy of language resourcedness by Simons et al. (2022) from Bird (2026).

At the lowest tier, the *Left-Behinds* (Class 0) represent digitally isolated languages lacking both labelled and unlabelled data, a category overlapping with Simons et al. (2022)'s *Still* layer which contains the vast majority of the world's languages (3,277) that remain completely unsupported. The *Scraping-Bys* (Class 1) possess limited unlabelled data but require structured, community-driven efforts to build labelled corpora, correlating with the *Emerging* (3,252 languages) and *Ascending* (456 languages) tiers where partial digital support begins to materialise. The *Hopefuls* (Class 2) show early promise through active support communities and small, existing labelled datasets, whereas the *Rising Stars* (Class 3) benefit significantly from unsupervised pre-training due to a strong web presence, yet remain bottlenecked by a lack of labelled data. Finally, the *Underdogs* (Class 4) possess massive amounts of unlabelled data and dedicated research communities, positioning them just below the *Winners* (Class 5). These top-tier winners map directly to the pyramid's elite *Vital* (108 languages) and *Thriving* (32 languages) tiers, representing the dominant, exceptionally high-resource languages backed by extensive datasets and substantial industrial and government investments.

These taxonomies correctly show that many LRLs, despite being spoken by a considerable portion of the global population, frequently lack the essential digital resources and tools to fully benefit from advancements in NLP (Salammagari & Srivastava 2024: 245).[5] However, as Bird (2026) argues, traditional paradigms of resource evaluation often fail to capture the complex realities of marginalised languages by relying fundamentally on absolute data quantities.

Proposing a more nuanced taxonomy to define *language resourcedness*, Nigatu et al. (2024: 17755–17757) shift the focus from a purely data-centric view to a multi-dimensional framework of four interacting criteria:

(i) *Sociopolitical factors*: societal, economic, historical, and political forces (e.g. financial constraints of data curation, colonial influences that limit language use in mainstream media, government, and education);
(ii) *Human and digital resources*: the availability of expert support (such as native speakers, linguists and NLP researchers) and the level of digital presence, where a lack of access to digital devices prevents language communities from being represented in standard web crawls and scraping;
(iii) *Artifacts*: the physical or digital availability of language tools, divided into linguistic knowledge (e.g. descriptions, orthographic standardisation), data (quantity, quality, benchmarks) and technology (from pre-processing tools to pre-trained multilingual LLMs);

[5] A distinction must be made between (i) strictly oral languages, (ii) languages with both spoken and written traditions, and (iii) extinct or historical languages preserved only in written form. Since current LLMs scale primarily on textual data, the focus rests on the challenges faced by languages that possess a written standard but lack high-quality textual assets required by current NLP technologies.

(iv) *Agency*: the role of community members in decision-making, as language technology must be actively designed by and for the communities to reflect their specific values.

In fact, LRLs are rarely just data-scarce. As Kaffee et al. (2023: 10:7) note, they are systematically "less studied, [...], less computerised, less privileged, less commonly taught, or low density, among other denomination."

The issue is that the current LLM pre-training paradigm disadvantages not only LRLs as defined by Kaffee et al. (2023), but also languages that, according to Nigatu et al. (2024)'s *resourcedness* framework, (i) face minimal political and economic constraints, (ii) boast millions of speakers and are academically studied, (iii) have existing corpora, and (iv) generate numerous resources daily. As Hämäläinen (2021: 1–2) affirms:

> At any case, I would never call my native language low-resourced. Why is this, you might ask? Calling Finnish low-resourced is denying the fact that we have our own TV shows, movies, music, theater plays, novels and other cultural products in Finnish. And I am not even talking about small numbers here. Besides, Finnish is a language of education, there is no level of schooling you could not complete entirely in Finnish, from preschool to defending your PhD. Yes, our language is small on a global scale, but we do have a whole bunch of resources we generate every day by communicating with each other!

It is the very architecture of contemporary AI systems, characterised by massive, web-scale pre-training on predominantly English data which creates a systemic bias that disadvantages nearly all other languages. Nonetheless, this issue goes beyond mere data volume, encompassing data composition, underlying architectural assumptions, and the political and economic forces driving model development. Consequently, a tiered system of linguistic performance emerges. Even languages with substantial resources show biases and deficiencies, and LRLs face exponentially increasing marginalisation. This occurs even when LRLs possess extensive digital and non-digital documentation and active communities of practice, all of which are disregarded by this exclusively data-driven methodology (cf. Hämäläinen et al. 2021).

3 Linguistic bias within the current AI paradigm

As we have already stated, the pre-training data for foundational LLMs exhibits a significant linguistic imbalance, heavily favouring English. For instance, approximately 93% of the data used for the now-dated GPT-3 model was in English (Brown et al. 2020). A similar pattern is evident in LLaMa 3.1, which allocated only 8% of this to multilingual content, despite being pre-trained on a massive 15 trillion-token dataset (cf. Grattafiori et al. 2024). This 8% multilingual portion spans 176 languages, meaning the remaining 92% of the dataset, primarily English, overwhelmingly outweighs the representation of all other languages combined

(ibid.). Consequently, many widely spoken languages (possessing an established written tradition; cf. Footnote 5) are severely underrepresented in the training corpus. The data distribution in LLMs like GPT-3 (Johnson et al. 2026) often shows a notable disparity when compared to the global number of speakers for various languages (Eberhard et al. 2025). For example, despite having 558 million speakers, Spanish only constitutes 0.8% of the data. Similarly, French, with 312 million speakers, accounts for 1.8%; German, with 134 million speakers, accounts for 1.5%; and Italian, with 66 million speakers, accounts for just 0.6%.

It is, however, important to acknowledge that the multilingual capabilities of LLMs are not inherently detrimental. Cross-lingual transfer can yield significant positive effects, particularly by improving performance for underrepresented languages when compared to purely monolingual systems (Conneau et al. 2020; Wu & Dredze 2020). Shared architectural representations allow lower-resource languages to leverage the syntactic and semantic knowledge of higher-resource languages. For instance, recent empirical work on endangered language pairs, such as the application of contextual embeddings for bilingual sentence mining in Upper and Lower Sorbian (Okabe & Fraser 2025) or speech recognition for Ryukyuan and Ainu, two dialects of Japan (Matsuura et al. 2026), demonstrates that targeted pre-training can facilitate positive cross-lingual transfer within related linguistic families.

However, while these architectures enable basic multilingual generation, recent efforts, such as NLLB-200 (No Languages Left Behind; cf. Costa-jussà et al. 2022), Seamless (cf. Barrault et al. 2023) and Omnilingual SONAR (Janeiro et al. 2026) by Meta AI, indicate that relying solely on cross-lingual transfer is insufficient. Empirical evidence from machine translation tasks involving over 1,500 languages in the BIBLE dataset within a state-of-the-art foundational embedding space trained on 200 languages reveals that translation quality fails to reach a "passable" threshold for the vast majority of languages, even when translating into English (Janeiro et al. 2026). Furthermore, Chang et al. (2026) reveals that multilingual LLMs struggle so significantly with basic next-token prediction in low-resource contexts that they are frequently surpassed by simple baseline bigrams, traditional statistical models that predict a word's probability based on the raw count of the single word immediately preceding it. Notably, compact monolingual models (125M parameters) for 350 different languages trained on 1 GB of data or less can outperform larger multilingual LLMs in perplexity and grammatical accuracy (Chang et al. 2026). This suggests that cross-lingual transfer should primarily be viewed as a fall-back when developing language-specific high-quality, annotated datasets is not feasible.

In addition, multilingual datasets inherently reflect the linguistic patterns, stereotypes, and cultural attitudes of their source societies (Johnson et al. 2026), which inevitably leads to the propagation of intrinsic biases related to ethnicity (Mitchell et al. 2025), gender (Savoldi et al. 2025), and socio-economic status (cf. Ferrara 2023: 4). This phenomenon is specifically referred to as *data bias*:

distortions resulting from the unequal distribution of information within the foundational data which proves that cross-lingual transfer, while technically functional, often erodes the integrity of marginalised languages (Laskar et al. 2026).

Recent research increasingly examines the linguistic influence of English on the multilingual outputs of various LLMs. Several studies focusing on Italian document this trend, revealing deviations from standard Italian norms. These include, albeit not comprehensively, variations in syntax and information structure (Cicero 2023, 2025; De Cesare 2025, 2026), orthography (Tavosanis 2024; e.g., *clarezza* vs. *chiarezza* 'clarity'), calques and borrowings (De Cesare 2024, 2026; Tavosanis 2024; cf. Example 1), the use of commas with sentence-initial adverbs and adverbials, and in syndetic constructions (De Cesare 2023, 2026; cf. Examples 2 and 3, respectively), a preference for pre-verbal subject positioning (Antonelli 2024, 2025), occasional marked sentence structures[6] (Tavosanis 2025), and the over-codification of highly accessible discourse referents (De Cesare 2025; cf. Example 4)

(1) In italiano, spesso si *omittano* i pronomi soggetto poiché la coniugazione verbale fornisce già informazioni sulla persona.
'In Italian, subject pronouns are often omitted because the verb conjugation already provides information about the person.' [GPT-3.5, De Cesare 2024: 81; my transl.]

(2) *Inoltre,* la Hack ha ricevuto numerosi premi [...].
'In addition, Hack has received numerous awards.' [GPT-3.5, De Cesare 2023: 204]

(3) Eleonora Duse era nota per le sue intense interpretazioni dei drammi di Gabriele d'Annunzio, Anton Cechov*, e* Henrik Ibsen.
'Eleonora Duse was known for her intense performances of the plays by Gabriele d'Annunzio, Anton Chekhov, and Henrik Ibsen.' [GPT-4, De Cesare 2024: 86; my transl.]

(4) *Eleonora Duse* nacque il 3 ottobre 1858 a Vigevano, in Italia, in una famiglia di attori. Spesso riconosciuta come una pioniera del metodo di recitazione, *Eleonora Duse* è notoriamente nota per il suo contributo unico e affascinante all'industria del teatro.

[6] Tavosanis (2025) identifies 18 total deviations within a 7,039-word corpus, averaging one error every 391 tokens. The taxonomy reveals that the most frequent disruptions involve coordinate constructions, specifically the heterogeneity of coordinated elements (5 occurrences) and faulty links in correlative pairs (1 occurrence). These are accompanied by the non-explicitation of necessary structural elements (3 occurrences), alongside recurring errors in number agreement, preposition selection, and article usage (2 occurrences each).

> ‘Eleonora Duse was born on October 3, 1858, in Vigevano, Italy, into a family of actors. Often recognized as a pioneer of the acting method, Eleonora Duse is famously known [famous] for her unique and fascinating contributions to the theatre industry.’ [GPT-4, De Cesare 2025: 46]

Taken together, these linguistic deviations unequivocally demonstrate that the textual competence of LLMs in Italian is systematically conditioned by cross-lingual interference from English. While the resulting prose may appear superficially fluent, it violates the macro-rules of structure, cohesion and style inherent to Italian.

German AI-generated texts, similarly, exhibit a noticeable English influence, particularly concerning reference management and agent realisation (Matthies 2025). This cross-lingual interference manifests as an overproduction of the impersonal agent (*unpersönliches Agens*), deviating from traditional German academic prose that typically favours agent declassification (*Agensdemotion*) via the passive voice and nominalisations (cf Example 5).

(5) *Dieses Kapitel untersucht* die Charakteristika […] der Relativsätze im Althochdeutschen und *bietet* Einblicke in ihre Funktion und Verwendung in verschiedenen Textsorten.
‘This chapter examines the characteristics [...] of relative clauses in Old High German and provides insights into their function and usage in various text types.’ [GPT-4, Matthies 2025: 104; my transl.]

Furthermore, syntax in German AI-generated texts displays a pronounced inclination toward SVO word order and reduced syntactic variability compared to texts produced by humans. According to Valentinelli (2026), AI-generated German texts exhibit a notably higher frequency of the Subject-Verb-Object (SVO) word order (cf. Example 6), constituting 38.3% to 83.1% of sentences, significantly surpassing the mean of 48.2% observed in a comparable human-authored reference corpus (Valentinelli 2026: 14–18).[7]

(6) *Die EU* muss jetzt ihre Politik und ihre Institutionen an die neue Realität anpassen und sicherstellen, dass sie weiterhin effektiv und relevant bleibt.
‘The EU must now adapt its policies and its institutions to the new reality and ensure that it remains effective and relevant.’ [llama-3.1-405b-instruct, Valentinelli 2026: 19]

Evidently, while the subject remains the most frequent initial element, other constituents follow in descending order of occurrence: adverbs (29.7% in the H-

[7] As noted by Valentinelli (2026), this structure effectively mirrors the fixed Subject-Verb-Object (SVO) order typical of English, despite the inherent V2 (Verb Second) nature of German syntax.

corpus versus 24.4% in the AI-corpus), direct objects (12.9% versus 0.5%), conjunctions (6.7% versus 3.3%), predicates (2.1% versus 1.3%), and, finally, whole sentences (1.7% versus 2.7%) (Valentinelli 2026: 18–29).

The examples presented so far, while not uniquely English linguistic features, clearly stem from an imbalanced pre-training dataset. As anticipated, these cross-lingual transfers are even more pronounced for LRLs. To illustrate this, let us now consider Example (7), which involves Cimbrian. As an endangered Upper German variety spoken in the Trentino-South Tyrol region of Italy (cf. Bidese 2021), Cimbrian is classified as an LRL according to Nigatu et al. (2024).

(7) In Lusern hot si am Sundàg gòte wàlo gsejn. Di Leit hoben gòte g'hòt unt di stimmen zò gòpe g'lòsn. Di neje president vo di Gemàinde isch Maria Toller, a frau mit fil g'muat unt g'spür fir di zimbarische sachn.
'On Sunday, elections took place in Luserna. The people turned out and cast their votes. The new president of the municipality is Maria Toller, a woman with a lot of heart and a strong sensibility for Cimbrian affairs.' [GPT-5; my transl.]

The challenge in generating Cimbrian text is not primarily the influence of English, but rather the competition from other languages with more available data. Both lexically (cf. Nicolussi Golo & Nicolussi 2013) and syntactically (cf. Bidese, Padovan & Tomaselli 2020), the sentences are not acceptable. For instance, Cimbrian neither exhibits the capitalisation of nouns like German does (cf. *Sundàg*, *Leit*, *Gemàinde*), nor Standard German V2 syntax (cf. In Lusern *hot* si am Sundàg gòte wàlo *gsejn*.). Simultaneously, the LLM appears to compensate for data scarcity by incorporating lexical transfers from various Germanic varieties. This includes words borrowed from Standard German (e.g., *am*, *stimmen*, *frau*, *mit*, *zimbarische*), from German dialects (e.g., *Leit, unt, Gemàinde, fil, fir*), and possibly even *Sundàg* from Norwegian Nynorsk. The core issue is the generation of Cimbrian output competing against the dominance of better-resourced languages.

4 Conclusion: A tripartite framework for LLM resourcedness

Previous chapters established the historical context of language classification in NLP, charting the shift from early definitions of *high-resource* versus *low-resource* languages to the modern era of pre-trained LLMs. Furthermore, we demonstrated that the current distribution of languages in AI training data creates significant linguistic, cultural, demographic, ideological, and political biases.

Previous taxonomies by Joshi et al. (2020) and Simons et al. (2022) proved inadequate by overlooking more nuanced and relevant criteria that have historically defined LRLs, whilst the *resourcedness* framework by Nigatu et al. (2024) neglects the specific composition of LLM pre-training datasets. The comparison in Table 1

highlights how established taxonomies often fail to account for AI-specific data imbalances.

Table 1: Dataset representation vs. taxonomy profiles for the top five languages in GPT-3.

Languages	%	Joshi et al. (2020)	Simons et al. (2022)	Nigatu et al. (2024)
English	93%			
French	1.8%			
German	1.5%	Winner	Thriving	High-resourced
Spanish	0.8%			
Italian	0.6%			

At the same time, while the simple dichotomy of English as the sole high-resource language and others as low-resource may hold a kernel of truth in the current AI paradigm, it is ultimately too simplistic (cf. Hämäläinen et al. 2021). Thus, defining the *status quo* of languages in the age of AI requires careful consideration.

To address the disparities in linguistic data distribution within LLMs, we propose adapting the *resourcedness* framework introduced by Nigatu et al. (2024). As we have already stated in Chapter 2, a language's resourcedness is evaluated through four exogenous criteria: (i) sociopolitical factors, (ii) human and digital resources, (iii) artifacts, and (iv) agency. While satisfying these criteria ensures the real-world vitality of a language, driving the creation of corpora and cultivating political and financial support, this real-world robustness does not linearly translate to equitable representation within AI systems. Instead, in the context of LLM pre-training, data volume and linguistic distribution dictates output quality and mitigates bias. Therefore, we argue that real-world *resourcedness* alone is insufficient to define a language as *high-resource* within the AI paradigm. For this reason, we also propose an endogenous criterion: a relative and model-agnostic quantitative threshold. To be considered truly *high-resource*, a language must hold a *dominant* or *co-dominant* statistical share in an LLM's pre-training dataset (e.g., reaching a proportional baseline, such as an equal share in a perfectly balanced multilingual LLM). This dual approach combining real-world criteria with an LLM-specific data threshold yields a new tripartite terminology:

(i) *High-resource*: Languages that meet all four *resourcedness* criteria and achieve *dominant* or *co-dominant* data representation.
(ii) *Few-resource*: Languages that possess strong real-world infrastructure, but fail to reach the proportional threshold within the LLM pre-training data.
(iii) *Low-resource*: Languages that are missing one or more real-world *resourcedness* criteria and inherently fall below the required algorithmic threshold.

Table (2) summarises the criteria and the new terminology.

Table 2: Exogenous and endogenous factors in defining *high-*, *few-*, and *low-resource languages*.

	Exogenous (Nigatu et al. 2024)				Endogenous
Terms	Socio-politics	Resources	Artifacts	Agency	Representativeness
High-resource	✓	✓	✓	✓	Dominant / Co-dominant
Few-resource	✓	✓	✓	✓	Marginalised
Low-resource	`Missing ≥1 criterion`				Marginalised

To illustrate the strengths and weaknesses of this terminology, we can apply this framework to the GPT-3 pre-training dataset (Brown et al. 2020). Under this proposal, English satisfies both the four exogenous factors and the endogenous factor, firmly establishing it as a *high-resource* language. Conversely, German and Italian meet the same four exogenous criteria for *resourcedness* as English, but because they fail to reach the proportional data threshold, they are more accurately categorised as *few-resource*. Cimbrian presents a distinctly different profile: while it benefits from a relatively strong exogenous support (a limited but vital speaker community, active scholarly attention, and socio-political backing from the Province of Trento), it severely lacks the digital footprint and technological infrastructure required to compile viable training data for LLMs, and, consequently, is defined as *low-resource*.

However, these definitions are not static. The very descriptors of languages vary from one LLM to another. Let us now take into consideration a bilingual LLM: Minerva LLM. According to Orlando et al. (2024), the Minerva LLM was trained on a total of 2.48 trillion tokens, consisting of 45,97% (1.14T tokens) of English data, 45,97% (1.14T tokens) of Italian data, and 8,06% (200B tokens) of programming code. Excluding code, the pre-training dataset exhibits a precise 50-50 linguistic distribution, therefore, both languages meet the *co-dominant* criterion and are considered *high-resource*.

Finally, LLMs such as BLOOM are trained with an objective that actively attempts to reduce the extreme prominence of English. Developed and released through a collaboration of hundreds of researchers (cf. Laurençon et al. 2022), BLOOM is a 176 billion parameter LLM trained on a diverse set of 46 natural and 13 programming languages. In the context of BLOOM's training data, English constitutes approximately 30% (for the full linguistic breakdown, see BigScience Workshop 2023: 8). While this is a significantly lower share compared to traditional LLMs, English remains the statistical majority by a wide margin. Therefore, under our proposed framework, English remains *high-resource*. Conversely, languages like Simplified Chinese (16.21%), French (12.93%), and Spanish (10.88%) meet all real-world exogenous criteria but fail to reach the dominant threshold set by English. Consequently, despite their massive absolute data volumes in the BLOOM dataset, they are accurately classified as *few-resource* within this specific architecture. This highlights the core argument of this paper: the critical factor is

not just the sheer volume of data, but rather the relative algorithmic distribution across the LLM's architecture.

Admittedly, the primary limitation of this framework hinges on a variable currently in short supply: data transparency. Because our endogenous criterion requires precise knowledge of linguistic distribution within pre-training datasets, the industry-wide shift towards proprietary *black boxes* poses a practical challenge. Leading AI companies have largely ceased disclosing their pre-training data composition, leaving researchers to rely on outdated metrics. However, the inability to audit a proprietary LLM does not invalidate our terminology; rather, it emphasises an urgent need to adopt the transparency standards already championed by open-source LLMs.

Nonetheless, while exact percentages are currently obscured, the structural imbalances of pre-training datasets remain largely undisputed. According to Villalobos et al. (2024), current trajectories in LLM development suggest that pre-training datasets will likely match the total available stock of public human-generated text between 2026 and 2032. Within this digital landscape, metrics from CommonCrawl (2026) highlight a significant linguistic concentration, with English representing around 41% of the data in such sources followed by a sharp drop-off to Russian (6.5%), German (5.9%), Japanese (5.1%), and Chinese (4.8%). Thus, it is highly probable that the linguistic composition of state-of-the-art LLMs has not fundamentally shifted over the years, cementing English as the single, unchallenged high-resource language within the current AI paradigm.

Essentially, adopting this tripartite framework (divided in *high-*, *few-* and *low-resource languages*) is fundamental for the critical evaluation of AI systems. By separating a language's actual sociopolitical status from its digital presence within LLMs' datasets, we break away from an obsolete binary classification. This prevents the false equivalence of labelling languages like Italian or German as *high-resource* when they are functionally marginalised within an LLM's architecture, while simultaneously preventing them from being dismissed as *low-resource* despite their robust societal infrastructure. By accurately categorising these structurally established but algorithmically underrepresented languages as *few-resource*, this terminology equips researchers and scholars with a precise, quantifiable tool.

In conclusion, while this framework provides a necessary theoretical foundation for redefining language resourcedness, its primary limitation remains its conceptual nature. Future research must prioritise the empirical validation of these categories. This requires testing the proposed framework across a broader typological range of languages, diverse open-source and proprietary LLMs, and specific downstream NLP tasks to quantify the practical impact of algorithmic marginalisation and test the efficacy of the proposed categories.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

References

Antonelli, Giuseppe. 2025. Storia brevissima (ma molto intensa) dell'IA-taliano. *Lingue E Culture Dei Media* 9(1), 4–50. https://doi.org/10.54103/2532-1803/29341

Antonelli, Giuseppe. 2024, L'italiano «autentico» dell'intelligenza artificiale. *AggiornaMenti. Rivista dell'Associazione di Docenti di Italiano in Germania* XV(15), 6–12. https://adi-germania.org/aggiornamenti-25/ (last accessed 25/06/2026).

Barrault, Loïc & Chung, Yu-An & Coria Meglioli, Mariano & Dale, David & Dong, Ning & Duppenthaler, Mark & Duquenne, Paul-Ambroise & Ellis, Brian & Elsahar, Hady & Haaheim, Justin et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv*. https://doi.org/10.48550/arXiv.2312.05187

Bidese, Ermenegildo. 2021. Introducing Cimbrian. The main linguistic features of a German(ic) language in Italy. *Energeia* 46. 19–62.

Bidese, Ermenegildo & Padovan, Andrea & Tomaselli, Alessandra. 2020. Rethinking V2 and nominative case assignment: new insights from a Germanic variety in Northern Italy. In Woods, Rebecca & Wolfe, Sam (eds). *Rethinking Verb Second*, 1–23. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198844303.001.0001

Bird, Steven. 2026. Towards a general theory of linguistic diversity. *Proceedings of the SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026*, 169–182. European Language Resources Association (ELRA).

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda et al. 2020. Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Chang, Tyler A. & Arnett, Catherine & Tu, Zhuowen & Bergen, Benjamin. 2026. Goldfish: Monolingual language models for 350 languages. *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC 2026), 3750–3781. European Language Resources Association (ELRA). https://doi.org/10.63317/5ceec3hhv4d5

Cicero, Francesco. 2023. L'italiano delle intelligenze artificiali generative. *Italiano LinguaDue* 15(2), 733–761. https://doi.org/10.54103/2037-3597/21990

Cicero, Francesco. 2025. Esercizi di stile. La narrativa per l'infanzia delle intelligenze artificiali. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 2(2). https://doi.org/10.62408/ai-ling.v2i2.19

CommonCrawl. 2026. Statistics of Common Crawl monthly archives: Distribution of languages. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html (last accessed 01/03/2026).

Conneau, Alexis & Khandelwal, Kartikay & Goyal, Naman & Chaudhary, Vishrav & Wenzek, Guillaume & Guzmán, Francisco & Grave, Edouard & Ott, Myle & Zettlemoyer, Luke & Stoyanov, Veselin. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2020.acl-main.747

Costa-jussà, Marta R. & Cross, James & Çelebi, Onur & Elbayad, Maha & Heafield, Kenneth & Heffernan, Kevin & Kalbassi, Elahe & Lam, Janice & Licht, Daniel & Maillard, Jean et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv*. https://doi.org/10.48550/arXiv.2207.04672

De Cesare, Anna-Maria. 2023. Assessing the quality of ChatGPT's generated output in light of human-written texts: A corpus study based on textual parameters. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos* 10, 179–210. https://revistas.uam.es/chimera/article/view/17979 (last accessed on 01/03/2026).

De Cesare, Anna-Maria. 2024. Nuove dinamiche di contatto linguistico. Le "impronte digitali" dell'inglese nell'italiano generato da LLM. *Lid'O. Lingua Italiana d'Oggi* XXI, 67–91.

De Cesare, Anna-Maria. 2025. LLM referential chain generation: A qualitative case study based on Italian biographies produced by GPT-4. *Linguistik Online* 136(4), 25–52. https://doi.org/10.13092/8tppdv47

De Cesare, Anna-Maria. 2026. *L'italiano sintetico dell'intelligenza artificiale generativa*. Firenze: Cesati.

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Eberhard, David M. & Simons, Gary F. & Fennig, Charles D. 2025. *Ethnologue: Languages of the World*. 28th ed. Dallas: SIL International. https://www.ethnologue.com (last accessed on 01/03/2026).

Ferrara, Emilio. 2023. Should ChatGPT be biased? Challenges and risks of bias in Large Language Models. *First Monday* 28(11), 1–39. https://doi.org/10.5210/fm.v28i11.13346

Grattafiori, Aaron & Dubey, Abhimanyu & Jauhri, Abhinav & Pandey, Abhinav & Kadian, Abhishek & Al-Dahle, Ahmad & Letman, Aiesha & Mathur, Akhil &

Schelten, Alan & Vaughan, Alex et al. 2024. The Llama 3 herd of models. *arXiv*. https://doi.org/10.48550/arXiv.2407.21783

Hämäläinen, Mika. 2021. Endangered languages are not low-resourced! In Hämäläinen, Mika & Partanen, Niko & Alnajjar, Khalid (eds.), *Multilingual Facilitation*, 1–11. University of Helsinki. https://doi.org/10.31885/9789515150257.1

Janeiro, João Maria & Cabot, Pere-Lluís Huguet & Tsiamas, Ioannis & Meng, Yen & Iyer, Vivek & Ramírez, Guillem & Barrault, Loïc. 2026. Omnilingual SONAR: Cross-lingual and cross-modal sentence embeddings bridging massively multilingual text and speech. *arXiv*. https://doi.org/10.48550/arXiv.2603.16606

Johnson, Rebecca & Dias Duran, Leslye Denisse & Panai, Enrico & Menéndez González, Natalia & Pistilli, Giada & Kalpokienė, Julija & Bertulfo, Donald Jay. 2026. The ghost in the machine speaks with an American accent: cultural value drift in early GPT-3 and the case for pluralist evaluation of generative AI. *AI Ethics* 6 (212). https://doi.org/10.1007/s43681-026-01038-x

Joshi, Pratik & Santy, Sebastin & Budhiraja, Amar & Bali, Kalika & Choudhury, Monojit. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2020.acl-main.560

Jones, Karen Sparck. 1994. Natural Language Processing: A historical review. In Zampolli, Antonio & Calzolari, Nicoletta & Palmer, Martha (eds), *Current Issues in Computational Linguistics: In Honour of Don Walker*, 3–16. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-0-585-35958-8_1

Kaffee, Lucie-Aimée & Biswas, Russa & Keet, C. Maria & Vakaj, Edlira Kalemi & de Melo, Gerard. 2023. Multilingual knowledge graphs and low-resource languages: A review. *Transactions on Graph Data and Knowledge* 1(1), 10:1–10:19. https://doi.org/10.4230/TGDK.1.1.10

Laskar, Md Tahmid Rahman & Islam, Mohammed Saidul & Nayeem, Mir Tafseer & Bhuiyan, Amran, Rahman, Mizanur, Joty, Shafiq, Hoque, Enamul & Huang, Jimmy. 2026. Lost in Translation: Do LVLM judges generalize across languages? *arXiv*. Accepted at the Association for Computational Linguistics (ACL) 2026 Findings. https://doi.org/10.48550/arXiv.2604.19405

Laurençon, Hugo & Saulnier, Lucile & Wang, Thomas & Akiki, Christopher & Villanova del Moral, Albert & Le Scao, Teven & Von Werra, Leandro & Mou, Chenghao & Ponferrada, Eduardo González & Nguyen, Huu et al. 2022. The BigScience Roots corpus: A 1.6TB composite multilingual dataset. *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022), 31809–31826. New Orleans, United States: Neural Information Processing Systems Foundation, Inc. https://openreview.net/forum?id=UoEw6KigkUn (last accessed 01/03/2026).

Le Scao, Teven & Fan, Angela & Akiki, Christopher & Pavlick, Ellie & Ilić, Suzana & Hesslow, Daniel & Castagné, Roman & Luccioni, Alexandra Sasha & Yvon, François & Gallé, Matthias et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv*. https://doi.org/10.48550/arXiv.2211.05100

Matsuura, Kohei & Ashihara, Takanori & Kawahara, Tatsuya. 2026. Adapting pretrained models to endangered languages in Japan: A comparative study on Ryukyuan and Ainu speech recognition. *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC 2026), 1455–1463. European Language Resources Association (ELRA). https://doi.org/10.63317/3iw5dymnwwbr

Matthies, Jochen. 2025. Agensdemotion und wissenschaftliche Verfasserreferenz im Kontext des KI-gestützten Schreibens – Eine Pilotstudie. In Lykhnenko, Olha & Matthies, Jochen & Roelcke, Thorsten & Teufele, Lisa (eds), *Deutsch als Fremdsprache – Mehrsprachigkeit, Fachkommunikation, Digitalisierung und Kultur: Internationale Nachwuchstagung 2024 an der Technischen Universität Berlin*, 91–114. Berlin: Springer. https://doi.org/10.1007/978-3-662-70351-9_5

Mitchell, Margaret & Attanasio, Giuseppe & Baldini, Ioana & Clinciu, Miruna & Clive, Jordan & Delobelle, Pieter & Dey, Manan & Hamilton, Sil & Dill, Timm & Doughman, Jad et al. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), 11995-12041. Association for Computational Linguistics (ACL).

Nicolussi Golo, Andrea & Nicolussi, Gisella & Panieri, Luca (ed.). 2013. *Zimbarbort. Börtarpuach Lusérnesch – Belesch / Belesch – Lusérnesch* (Dizionario del cimbro di Luserna). Luserna: Kulturinstitut Lusérn (Istituto Cimbro di Luserna).

Nigatu, Hellina Hailu & Tonja, Atnafu Lambebo & Rosman, Benjamin & Solorio, Thamar & Choudhury, Monojit. 2024. The Zeno's paradox of 'low-resource' languages. In Al-Onaizan, Yaser & Bansal, Mohit & Chen, Yun-Nung (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17753–17774. Miami, Florida, USA: Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2024.emnlp-main.983

Orlando, Riccardo & Moroni, Luca & Cabot, Pere-Lluís Huguet & Conia, Simone & Barba, Edoardo & Orlandini, Sergio & Fiameni, Giuseppe & Navigli, Roberto. 2024. Minerva LLMs: The first family of Large Language Models trained from scratch on Italian data. *Proceedings of the Tenth Italian Conference on Computational Linguistics* (CLiC-it 2024), 707–719. Pisa, Italy: CEUR Workshop Proceedings.

Okabe, Shu & Fraser, Alexander. 2025. Bilingual sentence mining for low-resource languages: a case study on Upper and Lower Sorbian. Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, 11–19. March 4-5, 2025, Association for Computational Linguistics (ACL). https://aclanthology.org/2025.computel-main.2/ (last accessed 20/06/2025)

Pakray, Partha & Gelbukh, Alexander & Bandyopadhyay, Sivaji. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing* 31(2), 183–197. https://doi.org/10.1017/nlp.2024.33

Pava, Juan N. & Meinhardt, Caroline & Zaman, Haifa Badi Uz & Friedman, Toni & Truong, Sang T. & Zhang, Daniel & Cryst, Elena & Marivate, Vukosi & Koyejo, Sanmi. 2025. *Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts*. Stanford Human-Centered Artificial Intelligence (HAI). https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts (last accessed 20/06/2026).

Radford, Alec & Narasimhan, Karthik & Salimans, Tim & Sutskever, Ilya. 2018. Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (last accessed 01/03/2026).

Radford, Alec & Wu, Jeffrey & Child, Rewon & Luan, David & Amodei, Dario & Sutskever, Ilya. 2019. Language models are unsupervised multitask learners. *OpenAI*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (last accessed 01/03/2026).

Salammagari, Abhi Ram Reddy & Srivastava, Gaurava. 2024. Advancing natural language understanding for low-resource languages: Current progress, applications, and challenges. *International Journal of Advanced Research in Engineering and Technology* (IJARET) 15(3), 244–255.

Savoldi, Beatrice & Bastings, Jasmijn & Bentivogli, Luisa & Vanmassenhove, Eva. 2025. A decade of gender bias in machine translation. *PATTERNS* 6(6). https://doi.org/10.1016/j.patter.2025.101257

Simons, Gary F. & Thomas, Abbey L. L. & White, Chad K. K. 2022. Assessing digital language support on a global scale. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli & Heng Ji et al. (eds). *Proceedings of the 29th International Conference on Computational Linguistics*, 4299–4305. International Committee on Computational Linguistics.

Tavosanis, Mirko. 2024. Valutare la qualità dei testi generati in lingua italiana. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 1(1). https://doi.org/10.62408/ai-ling.v1i1.14

Tavosanis, Mirko. 2025. Grammatica generata. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 2(2). https://doi.org/10.62408/ai-ling.v2i2.23

Valentinelli, Paolo. 2026. Tracing English interference in AI-generated German: An analysis of word order and syntactic fronting. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 5(1). https://doi.org/10.62408/ai-ling.v5i1.36

Villalobos, Pablo & Ho, Anson & Sevilla, Jaime & Besiroglu, Tamay & Heim, Lennart & Hobbhahn, Marius. 2024. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. *Proceedings of the 41st International Conference on Machine Learning* (ICML'24) 235, 49523–49544.

Wu, Shijie & Dredze, Mark. 2020. Are all languages created equal in multilingual BERT? *Proceedings of the 5th Workshop on Representation Learning for NLP*, 120–130. Association for Computational Linguistics (ACL).