Corpus textuels : enjeux et défis

Anna Pappa
Université Paris 8 - LIASD

Depuis les débuts de la linguistique computationnelle, les corpus textuels ont joué un rôle essentiel dans l'étude des phénomènes linguistiques[1]. À l'origine, les corpus se limitaient à des extraits d'œuvres littéraires ou à des transcriptions de discours, principalement utilisés pour les analyses lexicales et syntaxiques [33]. Avec les avancées en intelligence artificielle (IA), leur rôle a évolué : les corpus constituent la base d'apprentissage, pour des tâches spécifiques, comme la traduction automatique, l'analyse de sentiments ou la génération de texte. Des modèles comme GPT-3 ou BERT dépendent entièrement de la diversité et la qualité des données d'entraînement. Cette dépendance aux corpus soulève des défis : comment garantir leur diversité et leur représentativité à grande échelle ? Comment limiter les biais ? Comment assurer une annotation efficace ? Comment les adapter à des domaines spécialisés comme la médecine [20] ou le droit [12] ?

Ce travail explore ces questions en décrivant les méthodes de création et d'annotation des corpus ainsi que les défis techniques et éthiques liés à leur exploitation. Nous montrons pourquoi, malgré leurs limites, ces ressources restent fondamentales et comment elles peuvent être optimisées pour répondre aux exigences croissantes des modèles d'intelligence artificielle.

1. Nous ne faisons pas de distinction stricte entre *linguistique computationnelle*, qui s'intéresse davantage aux fondements théoriques, et *traitement automatique du langage* (TAL), qui se concentre sur la construction de systèmes pratiques. Tous deux reposent aujourd'hui sur des approches basées sur des ensembles des données (corpus), voir [27] et [19].

DOI:10.48556/SIF.1024.25.109

Évolution des corpus

L'étude des corpus[2] textuels a évolué avec les progrès technologiques en linguistique computationnelle et en IA.

Dans les années 1950 et 1960, les corpus étaient relativement « petits », comme le *Brown Corpus*[3], construits manuellement et utilisés pour des analyses linguistiques basiques. Dans les années 1980–90, ils sont devenus plus volumineux comme le *British National Corpus*[4] et ont inclus des données textuelles variées.

À partir des années 2010, avec l'émergence du *deep learning*, la nécessité de jeux de données massifs et diversifiés s'est imposée. Des corpus tels que *Common Crawl*[5] et *ImageNet*[6] ont vu le jour, fournissant des milliards d'exemples pour entraîner des modèles comme GPT-3 et BERT. Ces ressources ont permis aux systèmes IA d'atteindre des performances jamais vues, enrichies par une diversité linguistique et contextuelle sans précédent.

Les types de corpus

Les corpus peuvent être classés en fonction de leur taille, de leur domaine d'application et de leur spécialisation [39]. On distingue principalement trois catégories : les corpus généraux, les corpus spécialisés, et les corpus multilingues. Les premiers sont des corpus volumineux couvrant un large éventail de genres et de registres[7].

Les corpus spécialisés comme *PubMed*[8] (littérature biomédicale) et *CaseLaw*[9] (analyses juridiques) sont dédiés à des domaines spécifiques. Ils sont essentiels pour des tâches nécessitant précision et fiabilité, comme l'extraction de termes techniques ou la détection d'entités. Ces corpus spécialisés peuvent être monolingues, pour assurer une forte cohérence terminologique.

Les corpus multilingues contiennent des textes dans plusieurs langues et sont utilisés pour des applications comme la traduction automatique et la

2. Un corpus est un ensemble structuré de textes collectés pour l'analyse linguistique ou l'apprentissage automatique [3]. Lorsqu'il est annoté et adapté à une tâche spécifique, il devient un jeu de données (*dataset*).
3. Brown Corpus (1961): un des premiers corpus, composé d'un million de mots, utilisé pour l'analyse linguistique (`https://en.wikipedia.org/wiki/Brown_Corpus`).
4. British National Corpus (BNC) : 100 millions de mots (`https://www.english-corpora.org/bnc/`).
5. Common Crawl : un référentiel de données web accessibles librement (`https://commoncrawl.org/`).
6. ImageNet : base de données d'images hiérarchisées selon WordNet (`https://www.image-net.org/`).
7. Par exemple, *Common Crawl*, *Wikipedia* et *BookCorpus*, qui sont couramment utilisés pour étudier les phénomènes linguistiques, la traduction automatique et l'entraînement de modèles comme GPT et BERT.
8. `https://pubmed.ncbi.nlm.nih.gov/about/`.
9. `https://case.law/`.

modélisation sémantique. Des ressources comme *MOSAICo*[10] et *OPUS*[11] facilitent l'apprentissage des modèles de traduction automatique et l'étude des relations interlingues.

L'augmentation des données disponibles et la diversité des corpus apportent de nouveaux défis concernant leur collecte, structuration et annotation.

Constitution des corpus

L'explosion des données disponibles sur le Web a facilité la collecte de textes provenant de sources variées [22]. Le choix de sources dépend des objectifs visés : par exemple les blogs, forums, et réseaux sociaux alimentent des corpus destinés à analyser les tendances et les préférences des utilisateurs.

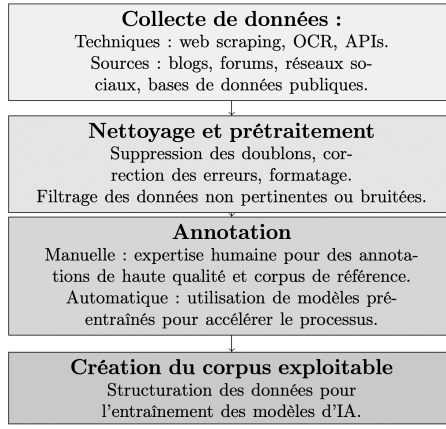

Fig. 1. Étapes principales de la création d'un corpus.

Les corpus peuvent être extraits de textes d'archives[12], de répertoires spécialisés comme *JSTOR*[13] pour les sciences sociales, ou être créés via *crowdsourcing*[14], une méthode qui fait appel à un grand nombre de contributeurs pour annoter les données. Pour les langues à faibles ressources, des initiatives comme le *Masakhane Corpus* [1] montrent l'importance de l'engagement des communautés locales pour la transcription et la traduction [29]. Les progrès

10. MOSAICo (`https://github.com/SapienzaNLP/mosaico`) est un corpus connectant des textes bruts à des bases de connaissances explicites.
11. OPUS est une collection de corpus parallèles pour la traduction automatique et l'étude des relations interlingues (`https://opus.nlpl.eu/`).
12. Par exemple le Projet Gutenberg pour les œuvres littéraires (`https://www.gutenberg.org/`).
13. `https://about.jstor.org/`.
14. Le *crowdsourcing* désigne une méthode de collecte et d'annotation de données via des plateformes en ligne [34].

technologiques comme le *web scraping* et la reconnaissance optique de caractères (OCR) permettent d'accéder rapidement à des volumes massifs d'information[3], mais ils présentent des défis : les données collectées peuvent être bruyantes, nécessitant des stratégies de filtrage et de nettoyage [7], tandis que l'OCR doit surmonter les lacunes dues aux textes mal numérisés [15]. Le diagramme de la figure 1 résume les étapes principales de la création d'un corpus.

Corpus et modèles d'IA : quantité, qualité et biais

L'évolution des modèles d'IA repose sur une expansion massive des corpus d'entraînement. Chaque nouvelle génération s'est appuyée sur des jeux de données de plus en plus vastes et diversifiés. Par exemple, OpenAI a multiplié par dix la taille des corpus entre GPT-1 et GPT-2, une tendance qui s'est poursuivie avec GPT-3 et GPT-4 (cf. tableau 1). Les modèles de langage sont entraînés sur des centaines à des milliers de milliards de tokens, provenant de sources variées : pages web, livres numérisés, bases de données spécialisées, médias sociaux et code informatique.

Cependant, l'usage de corpus massifs soulève une question essentielle : existe-t-il un volume optimal de données permettant d'améliorer les performances sans gaspiller inutilement des ressources ?

La question de la quantité optimale de données

Les lois d'échelle (*scaling laws* [21]) suggèrent que l'augmentation simultanée du nombre de paramètres, de la quantité de données et des ressources de calcul améliore systématiquement la qualité des modèles. Le modèle *Chinchilla* [18], développé par *DeepMind*[15], a remis en question cette approche. En effet, *Chinchilla* a surpassé des modèles plus grands mais sous-exploités, démontrant ainsi qu'un équilibre précis entre la capacité du modèle et la quantité de données est essentiel. Cette observation est également soutenue par d'autres études [17], qui soulignent qu'une simple augmentation du volume de données ne garantit pas toujours de meilleures performances.

Par ailleurs, d'autres recherches récentes [36] vont encore plus loin en remettant en cause l'idée qu'un modèle plus grand est toujours préférable. Elles montrent qu'un modèle plus petit mais mieux conçu, avec une architecture optimisée et un corpus soigneusement sélectionné, peut rivaliser avec des modèles massifs. En d'autres termes, la seule accumulation de données ne suffit pas : la qualité, la diversité et la représentativité des corpus jouent un rôle tout aussi déterminant que leur volume.

15. Entreprise spécialisée en IA appartenant à Google.

Modèle	Taille de paramètres (milliards)	Corpus (tokens) & Taille de données (To)	Sources principales	Langues
GPT-3 (OpenAI, 2020)	~175	~300 G (~45 To)	Common Crawl (60 %), WebText2 (22 %), Books1+2 (16 %), Wikipedia (3 %)	Majoritairement anglais
GPT-4 (OpenAI, 2023)	non divulgué	Estimé $\geq$ 10 To	Web + sources privées	Multilingue
LLaMA 1 (Meta, 2023)	7-65	~1.4 To	Common Crawl, GitHub, Wikipedia (20 langues), Books3, arXiv	91 % anglais, 9 % espagnol, français
LLaMA 2 (Meta, 2023)	7-70	~2 To	Données publiques filtrées	20+ langues, anglais dominant
BLOOM (BigScience, 2022)	176	366 G (~1.6 To)	ROOTS (presse, code, forums, etc.), 45 langues	46 langues, anglais ~30 % dont français, espagnol, arabe
PaLM 2 (Google, 2023)	340	~3.6 To	Web filtré, Wikipedia, livres, médias sociaux	Multilingue (100+ langues)
Mistral 7B (Mistral AI, 2023)	7	~2 To (estimé)	Web filtré, sources publiques	Majoritairement anglais
Chinchilla 70 (2023)	70	~1.4 To	Web filtré, livres, Wikipedia	Majoritairement anglais

Table 1. Corpus d'entraînement des principaux modèles de langage.

L'annotation des corpus : entre qualité et automatisation

L'augmentation des volumes de données a rendu le filtrage, le nettoyage et l'annotation indispensables. La qualité de l'annotation est un facteur déterminant pour l'efficacité des modèles de langage [26], mais elle pose des défis en termes

de coût, de temps et de fiabilité. Pour y répondre, plusieurs approches ont été développées, allant de l'apprentissage supervisé à l'automatisation partielle.

L'apprentissage supervisé, bien qu'essentiel, repose sur des annotations manuelles coûteuses et longues à produire. Des outils comme *Snorkel*[16] utilisent la supervision faible, génèrent des annotations préliminaires pour réduire le besoin d'intervention humaine.

Le *Cross-View Training* (CVT, voir [10]), exploite des données non annotées pour améliorer l'annotation automatique combinant annotations de qualité et prédictions sur de grands volumes. L'apprentissage actif (*active learning*) cible les échantillons les plus informatifs ou incertains pour une annotation manuelle, optimisant l'efficacité. Des outils comme *Prodigy*[17] intègrent cette technique, souvent combinée à d'autres méthodes [6].

L'annotation assistée par l'IA, via des modèles comme GPT-4, génère des annotations préliminaires avec précision, réduisant temps et coûts [11]. Cependant, une validation humaine reste nécessaire pour corriger les biais et erreurs, assurant un compromis optimal entre efficacité et précision.

En somme, l'hybridation des méthodes — supervision faible, assistance par LLM, apprentissage actif et validation humaine — permet de créer des corpus de haute qualité, essentiels pour des modèles d'IA performants et fiables.

Représentativité et biais dans les corpus

La représentativité des corpus est un enjeu à la fois technique et éthique qui affecte directement la qualité des modèles d'intelligence artificielle. Lorsqu'un corpus est dominé par certaines langues ou aspects culturels, il introduit des biais qui influencent les résultats des modèles entraînés [23]. Par exemple, une surreprésentation de l'anglais peut marginaliser d'autres langues et cultures, empêchant les modèles de bien fonctionner dans des contextes multilingues et diversifiés.

Pour faire face à ces biais, plusieurs approches ont été développées. L'*échantillonnage stratifié* permet d'équilibrer la représentation linguistique et culturelle, en garantissant la présence proportionnelle de différentes sources [8]. La *débiaisation*[18] *algorithmique*, consiste à corriger les biais présents dans les données en ajustant statistiquement leur importance [5]. FENEC est un exemple de corpus équilibré qui tente de garantir une représentation diversifiée des genres textuels [28]. Enfin, l'approche par *crowdsourcing* implique

16. Snorkel (`https://www.snorkel.org/`) : un outil puissant pour les tâches en TAL et la création de modèles d'apprentissage supervisé à partir de données faiblement étiquetées.
17. Prodigy (`https://prodi.gy/`) est un outil d'annotation extensible pour créer des systèmes d'IA personnalisés.
18. Le terme *débiaisation* est une traduction courante de l'anglais *debiaising*, utilisée dans la littérature en ligne. Bien que cette traduction ne soit pas encore standardisée en français, elle est fréquemment employée pour désigner les techniques visant à réduire les biais algorithmiques.

directement les communautés concernées et permet de constituer des corpus plus représentatifs des réalités linguistiques et culturelles locales [29].

Cependant, un défi supplémentaire se pose : l'impact environnemental des grands corpus. Leur collecte, stockage et traitement nécessitent des ressources énergétiques importantes, entraînant un coût écologique significatif. Trouver un équilibre entre diversité, représentativité et durabilité écologique est devenu essentiel pour garantir une exploitation responsable et efficace des corpus à l'ère de l'IA [32].

Impact environnemental et gestion durable des corpus

L'entraînement des modèles d'IA sur des corpus massifs engendre une consommation énergétique considérable, soulevant des préoccupations croissantes quant à leur impact écologique. La constitution et la gestion de ces ressources, mobilisent des infrastructures informatiques de grande envergure, bien avant même la phase d'apprentissage des modèles. Dès la collecte, le traitement des corpus passe par des opérations gourmandes en ressources : acquisition de données hétérogènes, nettoyage, annotation et extraction d'informations [35].

L'empreinte énergétique de grands corpus s'étend au-delà de leur préparation initiale. Leur stockage à long terme exige des infrastructures informatiques conséquentes, et leur enrichissement continu pour maintenir leur pertinence accroît cette demande énergétique. Par ailleurs, la réutilisation de ces corpus lors des phases de *fine-tuning* et d'adaptation des modèles nécessite des calculs intensifs, surtout lorsque ces modèles sont fréquemment mis à jour avec de nouvelles données. Cette expansion continue contribue à alourdir l'empreinte carbone des infrastructures informatiques. Par exemple, l'entraînement de modèles tels que BERT peut consommer autant d'énergie que plusieurs dizaines de foyers sur plusieurs années [35]. Le modèle multilingue BLOOM a lui aussi illustré l'ampleur de cette empreinte écologique, en générant environ 50 tonnes de CO_2 lors de son entraînement [24]. Face à ces défis, plusieurs stratégies émergent pour minimiser l'impact environnemental des corpus :

– optimisation des infrastructures : utilisation de supercalculateurs alimentés par des énergies renouvelables [2], et des systèmes de gestion avancés[19] ;
– réduction et compression des données : techniques comme la quantification et la compression réduisent la taille des corpus sans compromettre leur diversité ni leur représentativité [25], ainsi que l'utilisation de modèles allégés [35] ;

19. Comme Hadoop et Spark qui facilitent l'analyse et le traitement des grands corpus, tandis que des outils spécialisés comme *ELAN* (annotateur linguistique, `https://www.mpi.nl/corpus/html/elan/`) et *ANNIS* (plateforme web de recherche et visualisation pour corpus annotés, `https://corpus-tools.org/annis/`), assurent leur structuration et leur exploitation.

– sobriété numérique : privilégier des corpus plus compacts mais mieux annotés et équilibrés en accord avec les lois d'échelle, qui montrent qu'un rapport optimal entre la taille du modèle et la quantité de données d'entraînement permet de maximiser les performances tout en réduisant l'empreinte énergétique [18].

La gestion des corpus doit évoluer vers des pratiques plus durables. La combinaison de stratégies d'échantillonnage intelligent, de techniques de compression et de réduction du stockage redondant pourrait constituer une alternative viable pour concilier l'efficacité des modèles et la préservation des ressources énergétiques.

Défis juridiques, confidentialité et données synthétiques

Outre ces considérations techniques, l'exploitation de corpus à grande échelle soulève d'importantes questions au niveau juridique. De nombreux corpus spécialisés (médicaux, juridiques, journalistiques) sont protégés par des droits de propriété intellectuelle, limitant leur utilisation pour l'entraînement des modèles d'IA. Par exemple, *Meta* a été accusé d'avoir intégré des livres protégés par le droit d'auteur dans l'entraînement de LLaMA 2, notamment des documents issus de la bibliothèque pirate *Library Genesis* [20].

De même, *OpenAI* fait face à des poursuites judiciaires intentées par des auteurs et éditeurs, comme George R.R. Martin et John Grisham[21], pour avoir utilisé des œuvres sous copyright sans autorisation dans GPT-3 et GPT-4.

Pour répondre à ces problématiques, plusieurs solutions sont envisagées :

– utilisation de textes du domaine public (par exemple le *Projet Gutenberg*) ou de corpus ouverts (par exemple *RedPajama*) ;
– négociation des licences spécifiques pour l'utilisation des textes protégés ;
– adaptation du *Fair Use*[22] pour permettre un accès limité aux œuvres sous copyright, sous réserve d'un usage strictement encadré.

Parallèlement, les données synthétiques peuvent être utilisées pour compléter ou remplacer des jeux de données réels, en particulier quand les données sont rares, déséquilibrées ou difficiles à collecter [30]. Ces données, générées artificiellement, permettent d'augmenter la taille des corpus tout en évitant les problèmes de droits d'auteur. Cependant, leur adoption fait face à de nombreuses critiques, notamment en ce qui concerne leur fiabilité, leur potentiel à reproduire des biais existants, et les risques liés à la protection des données personnelles [14]. Pour

20. Voir l'article de Reuters (2025).
21. Article de Reuters (2023).
22. Le *Fair Use* (utilisation équitable) est une doctrine juridique américaine permettant une utilisation limitée d'œuvres protégées sans autorisation sous certaines conditions. Un cas célèbre est celui de Google, qui a utilisé cette doctrine pour justifier la numérisation de millions de livres à des fin de recherche et d'indexation, *Authors Guild, Inc. v. Google, Inc.*

répondre à ces préoccupations, des protocoles rigoureux sont essentiels afin de garantir leur qualité, leur transparence et leur conformité aux réglementations en vigueur [9]. En somme, la constitution des corpus doit combiner des impératifs techniques, juridiques et éthiques, tout en explorant des approches alternatives pour répondre aux besoins croissants de l'IA.

Le corpus optimal à l'ère de l'IA générative

La constitution de corpus pour l'entraînement des modèles d'IA repose sur trois critères essentiels : la qualité, la représentativité et la taille des données [4]. Bien que chaque aspect soit souvent étudié individuellement, à l'ère des modèles génératifs comme GPT-4, une approche plus holistique est nécessaire. L'objectif est de définir un « corpus optimal » qui intègre et équilibre, de façon cohérente, les trois dimensions suivantes :

– la qualité ;
– la représentativité ;
– la taille.

Un corpus optimal doit tout d'abord garantir une qualité élevée, c'est-à-dire des données fiables, correctement annotées et exemptes de biais majeurs [5]. Ensuite, il doit être représentatif, reflétant la diversité linguistique, culturelle et thématique nécessaire pour assurer une bonne généralisation[23].

Enfin, la taille du corpus doit être adaptée : ni trop réduite pour éviter de limiter les performances, ni excessivement volumineuse, car une augmentation excessive des données ne garantit pas systématiquement de meilleurs résultats [18].

Cependant, les modèles de grande taille, bien qu'impressionnants, ne comprennent pas le langage. Ils apprennent à partir d'associations statistiques dérivées des corpus, ce qui peut entraîner des biais, des erreurs et parfois des hallucinations[24], même lors du préentraînement adaptatif[25] [16].

Pour répondre à ces défis, les corpus évoluent vers des formats enrichis : intégration de métadonnées, annotations détaillées et structures logiques pour mieux exploiter les connaissances [13]. Certains corpus récents intègrent même des faits issus de bases de connaissances, permettant aux modèles d'associer génération de texte et inférences logiques pour simuler un « raisonnement automatique »[26] (voir [31]). Cette capacité reste cependant une

23. Par exemple, un corpus médical doit inclure des textes scientifiques ainsi que des échanges réels entre professionnels de santé et patients, qui capturent mieux les situations pratiques et les besoins concrets [20].

24. Phénomène où un modèle génère des informations incorrectes ou inventées, souvent de manière convaincante.

25. Il s'agit d'un modèle initialement entraîné sur un large corpus et affiné avec des données plus spécifiques.

26. Les IA génératives comme *GPT-4* ou *Deepseek* peuvent résoudre des problèmes mathématiques ou expliquer des concepts complexes en détaillant chaque étape [37].

simulation statistique : elle améliore l'apparence logique des réponses sans conférer aux modèles une véritable capacité déductive comparable à celle d'un raisonnement humain complexe. Des travaux récents, tels que NELLIE [38], cherchent justement à introduire des mécanismes d'inférence explicite pour dépasser ces limitations.

En somme, le défi n'est plus seulement d'accumuler des données, mais de les sélectionner intelligemment, de les annoter avec précision et de les structurer pour garantir leur efficacité et leur adaptabilité aux modèles génératifs de nouvelle génération.

Conclusion

La constitution et l'exploitation des corpus textuels restent une tâche complexe, impliquant des enjeux techniques, éthiques et écologiques. Pour y répondre, une approche raisonnée et équilibrée s'impose : sélectionner soigneusement les données pour garantir leur qualité, veiller à leur diversité, et les structurer pour maximiser leur efficacité. Les corpus demeurent au cœur du développement des modèles d'IA, permettant à ces derniers d'être performants, fiables et respectueux des ressources environnementales et des exigences sociétales.

L'avenir des corpus réside dans la création d'écosystèmes modulaires et adaptatifs, où qualité, représentativité et taille sont équilibrées de manière dynamique, en réponse aux besoins évolutifs des modèles et des sociétés.

Références

[1] David Ifeoluwa Adelani and al. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. https://arxiv.org/abs/2210.12391.

[2] M. Amisse, M. Faur, L. Gonard, and A. Orcesi. 2024. *Promouvoir des modèles d'intelligence artificielle frugale pour et par les politiques publiques*. HAL. https://hal.science/hal-04510171v1/file/GAAP_rapport_IA_frugale.pdf.

[3] Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Demonstrations, 2006*, 11–16. https://aclanthology.org/W06-2202.

[4] Emily M. Bender and al. 2021. On the dangers of stochastic parrots: Can language models be too big? In (FAccT '21), 610–623. https://doi.org/10.1145/3442188.3445922.

[5] Tolga Bolukbasi and al. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NeurIPS 2016*, 4349–4357. https://papers.nips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[6] Marouaa Boudabous and Anna Pappa. 2023. Explicit aspect annotation via transfer and active learning. In *Procedia computer science*, 1124–1133.

[7] Maroua Boudabous and Anna Pappa. 2021. WebT-IDC: A web tool for intelligent dataset creation a use case for forums and blogs. In *Procedia computer science*. 1051–1060.

[8] Vaclav Brezina, Abi Hawtin, and Tony McEnery. 2021. *Text & Talk* 41, 5–6: 595–615. https://doi.org/doi:10.1515/text-2020-0052.

[9] Helena Canever. 2023. Réflexions sur la création de données synthétiques sûres et conformes à la réglementation. *1024* 22: 45–53. https://doi.org/10.48556/SIF.1024.22.45.

[10] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-supervised sequence modeling with cross-view training. https://arxiv.org/abs/1809.08370.

[11] Bosheng Ding and al. 2023. Is GPT-3 a good data annotator? In *ACL*, 11173–11195. https://doi.org/10.18653/v1/2023.acl-long.626.

[12] Chalkidis Ilias et al. 2020. LEGAL-BERT: The muppets straight out of law school. In *EMNLP 2020*, 2898–2904. https://aclanthology.org/2020.findings-emnlp.261/.

[13] Leo Gao and al. 2020. The pile: An 800GB dataset of diverse text for language modeling. https://arxiv.org/abs/2101.00027.

[14] Ian J. Goodfellow and al. 2014. Generative adversarial networks. https://arxiv.org/abs/1406.2661.

[15] Google Inc. 2017. Tesseract. (https://github.com/tesseract-ocr).

[16] Daniil Gurgurov and al. 2025. Small models, BIG impact: Efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages. https://arxiv.org/abs/2502.10140.

[17] Tom Henighan and al. 2020. Scaling laws for autoregressive generative modeling. https://arxiv.org/abs/2010.14701.

[18] Jordan Hoffmann and al. 2022. Training compute-optimal large language models. https://arxiv.org/abs/2203.15556.

[19] Daniel Jurafsky and James H. Martin. 2025. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models*. (Online manuscript released January 12, 2025). https://web.stanford.edu/~jurafsky/slp3/.

[20] Singhal K. and al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972: 172–180. https://doi.org/10.1038/s41586-023-06291-2.

[21] Jared Kaplan and al. 2020. Scaling laws for neural language models. https://arxiv.org/abs/2001.08361.

[22] Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics* 29, 3: 333–347.

[23] Geoffrey Leech. 2007. New resources, or just better old ones? The holy grail of representativeness. In *Corpus linguistics and the web*, M. Hundt, N. Nesselhauf and C. Biewer (eds.). Rodopi, 133–149.

[24] Alexandra Sasha Luccioni and al. 2022. Estimating the carbon footprint of BLOOM, a 176B parameter language model. In *EMNLP*.

[25] Aru Maekawa and al. 2023. Dataset distillation with attention labels for fine-tuning BERT. In *ACL*, 119–127. https://doi.org/10.18653/v1/2023.acl-short.12.

[26] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT.

[27] Tony McEnery and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

[28] Alice Millour and al. 2022. FENEC : Un corpus équilibré pour l'évaluation des entités nommées en français (FENEC: A balanced sample corpus for French named entity recognition). In *TALN*, 82–94. https://aclanthology.org/2022.jeptalnrecital-taln.8/.

[29] Wilhelmina Nekoto and al. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *EMNLP 2020*, 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195.

[30] Sergey I. Nikolenko. 2019. Synthetic data for deep learning. https://arxiv.org/abs/1909.11512.

[31] Fabio Petroni and al. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP*, 2463–2473. https://doi.org/10.18653/v1/D19-1250.

[32] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM* 63, 12: 54–63. https://doi.org/10.1145/3381831.

[33] John M. Sinclair. 1991. The automatic analysis of corpora. In *Corpus linguistics: Volume 2*. De Gruyter, 379–397.

[34] Rion Snow and al. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *EMNLP*, 254–263. https://aclanthology.org/D08-1027/.

[35] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *ACL*, 3645–3650. https://doi.org/10.18653/v1/P19-1355.

[36] Yi Tay and al. 2022. Scaling laws vs model architectures: How does inductive bias influence scaling? https://arxiv.org/abs/2207.10551.

[37] Jason Wei and al. 2023. Chain-of-thought prompting elicits reasoning in large language models. https://arxiv.org/abs/2201.11903.

[38] Nathaniel Wier, Peter Clark, and Benjamin van Durme. 2024.NELLIE: a neuro-symbolic inference engine for grounded, compositional, and explainable reasoning. Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2024/399.

[39] R. Xiao. 2010. Corpus creation. In *The handbook of natural language processing* (2nd ed.), N. Indurkhya and F. Damerau (eds.). CRC Press, London, 147–165.