Parole e lingua nell'universo GAFAM

Manuel Favaro

Università LUMSA
(manuel.favaro@lumsa.it)

Abstract

Il presente contributo si propone di analizzare i diversi servizi offerti da Google, Amazon e Microsoft per il machine learning e il trattamento automatico del linguaggio, con il proposito di esaminarne sia i pregi – immediatezza e facilità d'uso –, sia i difetti – asservimento a logiche commerciali, inefficienza di alcuni prodotti, specie per quelli che si pongono come punti di riferimento per professionisti e specialisti.

DOI: https://doi.org/10.58015/2036-2293/591

1. Introduzione

Dopo l'allentamento in Italia delle restrizioni dovute al Covid-19, uno dei simboli della riapertura e di un rinnovato fervore per l'arte è stata la mostra dedicata a Gustav Klimt (*Klimt. La Secessione e l'Italia*), tenutasi al Museo di Roma tra l'ottobre del 2021 e il marzo del 2022. Camminando tra le opere del genio austriaco e dei suoi colleghi viennesi, il visitatore si imbatteva nei *Quadri delle Facoltà*; o meglio: nella loro ricostruzione. Come sottolinea il sito della mostra[1], i tre dipinti, rappresentanti le allegorie *La Medicina*, *La Giurisprudenza* e *La Filosofia*, furono infatti commissionati a Klimt alla fine degli anni Dieci del Novecento per l'Aula Magna dell'Università di Vienna, ma vennero successivamente rifiutati perché ritenuti scandalosi. Andati infine perduti a causa di un incendio, avvenuto nel 1945, ciò che ne rimaneva erano fotografie in bianco e nero e articoli di giornale che ne descrivevano fattezze e colorazioni.

Solo nell'autunno del 2021, il Google Arts & Culture Lab Team, tra cui spicca la figura del programmatore Emil Wallner, avvalendosi della consulenza del dottor Franz Smola, uno dei maggiori esperti mondiali di Klimt, ha preso in mano il poco che era sopravvissuto e, tramite tecniche di apprendimento automatico, è riuscito nell'impresa di ricostruire quelli che presumibilmente erano i colori originali dei *Quadri delle Facoltà*.

Il restauro digitale è stato forse la sezione più interessante della ormai conclusa mostra, ed è al centro dell'esposizione virtuale *Klimt vs Klimt*, curata proprio da Google Arts & Culture. La mostra online raccoglie 120 capolavori dell'austriaco provenienti da 30 istituzioni culturali di tutto il

[1] http://www.museodiroma.it/it/mostra-evento/klimt-la-secessione-e-l-italia.

mondo[2], e offre all'utente un'esperienza del tutto inedita, grazie a una futuristica applicazione 3D di realtà aumentata che evidenzia ogni dettaglio delle immagini.

Dunque, non solo il lavoro di restauro, che per quanto digitale è assolutamente straordinario, ma anche la divulgazione internazionale della rediviva opera è gestita da Google. I quadri restaurati sono stati accompagnati – sia a Roma, sia nel sito della mostra virtuale – da un video che ritrae Wallner e Smola mentre raccontano la loro impresa, illustrando a migliaia di visitatori la potenza e soprattutto il fascino del machine learning, in questo caso applicato all'arte.

Per chi legge quella appena raccontata sembrerebbe una storia entusiasmante e a lieto fine: un colosso internazionale si mette a disposizione del popolo per salvaguardare il bene comune, riportando alla luce ciò che altrimenti non sarebbe stato possibile far risorgere. La domanda che però, a questo punto, sarebbe lecito porsi, è: solo Google e le altre aziende GAFAM, con la loro disponibilità economica e il loro strapotere, potevano arrivare a questo incredibile risultato? Oppure esistono altre istituzioni in grado di competere contro Big G & Co. senza ridurre l'apporto accademico alla mera consulenza di un singolo esperto?

2. Il trattamento automatico della lingua: Google, Amazon e Microsoft

2.1 Servizi efficaci per l'utente comune: Google Translate

Il machine learning riguarda una gamma di applicazioni che vedono tra i protagonisti il linguaggio. Mediante le tecniche di trattamento automatico, è ormai possibile fare (quasi) tutto: riconoscere e imitare la voce umana, tradurre testi automaticamente, rendere accessibili e interrogabili banche dati enormi tramite poche e semplici istruzioni da dare in pasto alla macchina. L'accesso ad alcune di queste risorse è tutt'altro che riservato agli specialisti, anzi: per tradurre un testo basta aprire Google Translate, copiarlo, incollarlo e aspettare che in pochi istanti il sistema neurale di traduzione automatica, il Google Neural Machine Translation (GNMT), converta il nostro testo nella lingua desiderata, in maniera se non perfetta almeno altamente accettabile, molto più di quanto non fosse il vecchio sistema probabilistico (cfr. Melvin et al. 2017)[3].

Un servizio, insomma, che qualunque persona, dal professionista all'utente comune, può usare ogni giorno per qualsiasi scopo. Ed è soprattutto quest'ultimo che in genere non conosce alternative, per il semplice fatto che non ne ha bisogno: perché dovrebbe cercare altri strumenti, quando ha gratuitamente a disposizione un servizio veloce e adatto alle sue esigenze quotidiane? Quando si cita Google Translate, si parla appunto informalmente di un servizio che nessuno, financo i suoi stessi sviluppatori, si sognerebbe di usare per tradurre testi importanti senza una doverosa e sistematica revisione manuale, perché si tratta di una risorsa messa a disposizione della comunità digitale senza particolari pretese scientifiche, e che invece si propone di fornire un supporto linguistico immediato, soprattutto per le lingue sconosciute all'utente. Sta poi allo stesso utilizzatore ricorrere a dizionari e grammatiche per confermare l'attendibilità della traduzione proposta da Google e apportare, dove servissero, le modifiche necessarie.

[2] https://artsandculture.google.com/project/klimt-vs-klimt.

[3] Per una buona panoramica sui nuovi sistemi neurali di traduzione automatica, si veda Stahlberg (2020).

2.2 Servizi inefficaci per l'utente specializzato: Google Ngram Viewer

Anche Google Ngram Viewer è un servizio gratuito e accessibile a tutti, ma il fatto di essere uno strumento molto meno diffuso rispetto a Google Translate può comportare un utilizzo meno consapevole in grado di fare maggiori danni, poiché i problemi strutturali della risorsa, come vedremo adesso, sono molto più profondi.

Lo strumento, infatti, permetterebbe – il condizionale è d'obbligo – di osservare l'evoluzione di una parola nel tempo mostrando all'utente la frequenza della stringa ricercata in un certo periodo, e lo fa utilizzando diversi corpora disponibili per alcune delle lingue più diffuse, tra cui l'italiano, che sono basati sull'enorme banca dati di Google Libri. Grazie a un sistema grafico accattivante, l'utente ha l'illusione di accedere subito alla storia di un vocabolo, e di conseguenza alla storia di un nome, di un oggetto, di una idea, di un concetto, come testimonia il nutrito numero di articoli accademici di qualsivoglia disciplina, reperibili nella famosa biblioteca digitale JSTOR[4], che citano Google Ngram Viewer, alcuni dei quali entusiasticamente[5], altri senza interrogarsi sulla sua effettiva affidabilità[6]; soprattutto quelli scritti a ridosso del 2010, anno in cui lo strumento veniva messo a disposizione del pubblico internazionale, il quale si illudeva che una risorsa di tal fatta permettesse di esplorare e trarre immediato vantaggio da un universo fino ad allora sconosciuto.

In un certo qual senso, ciò è vero: Google Ngram Viewer consente di avere una panoramica sulla distribuzione cronologica della parola o del sintagma ricercati. Tale panoramica però deve essere considerata preliminare allo studio effettivo. I motivi sono semplici: in primo luogo, non è dato conoscere la reale dimensione e composizione del corpus di una specifica lingua diversa dall'inglese, a meno di scaricarsi il dataset, quando disponibile, da risorse esterne come GitHub (cfr. § 3), così com'è impraticabile capire a quale dominio appartengano i testi, compromettendo a monte il criterio di rappresentatività, su cui si basano gran parte delle analisi sui corpora (cfr. Biber 1993). In secondo luogo, non è possibile accedere direttamente dalla risorsa ai passaggi contenenti la parola, e dunque non si possono osservare i vari contesti in cui essa si trova, procedimento essenziale ad analisi di tipo qualitativo: si può soltanto riflettere sulla distribuzione meramente cronologica delle parole, senza entrare nel merito di aspetti cogenti quali la tipologia testuale, il registro e altri fattori sia interni, sia esterni alla storia del vocabolo; va da sé che l'informazione fornita è quantomeno insufficiente per impostare ricerche che possano considerarsi valide. In terzo luogo, secondo quanto dichiarato dagli stessi sviluppatori[7], i 33 corpora, a rappresentanza delle lingue inglese (13), cinese (3), francese (3), tedesco (3), ebraico (3), spagnolo (3), russo (3) e italiano (2) sono stati rilasciati in tre periodi diversi (2009, 2012, 2019), un lungo periodo in cui la linguistica computazionale ha compiuto passi da gigante. Basti fare un esempio: nelle versioni del 2009, ancora oggi impiegate dalla risorsa, il processo di suddivisione delle forme presenti nelle unità di base del

[4] http://www.jstor.com.

[5] Per esempio, l'informatica umanistica Lauren Klein (2011: 37) parla di Google Ngram Viewer come uno strumento in grado di stimolare nuovi interessi, nuovi approfondimenti riguardo ai mutamenti intercorsi nel tempo tra lingua, cultura e letteratura.

[6] Si veda il contributo del teorico e critico letterario Walter Benn Michaels (2011: 310), il quale usa i risultati di Google Ngram Viewer a supporto di una sua ipotesi cronologica circa la diffusione del termine *postmodernismo*.

[7] https://books.google.com/ngrams/info.

testo digitale, ossia i *token*, nel gergo denominato appunto tokenizzazione, avveniva utilizzando lo spazio bianco come elemento separatore. Per intendersi: forme come *l'altro* o *esempio…* vengono considerati un token singolo anziché essere doverosamente scissi in token separati; risultati di questo tipo, a oggi, sarebbero inconcepibili, e infatti le versioni successive del 2012 e del 2019 usano metodi di segmentazione ibridi, basati su regole per il cinese, su sistemi statistici per tutte le altre lingue. Non è chiara neanche l'accuratezza dell'annotazione morfo-sintattica, ossia l'aggiunta di informazioni grammaticali ai singoli token, fondamentale nel processo di arricchimento linguistico del testo grezzo, già influenzata negativamente dai problemi di tokenizzazione: secondo quanto si legge sul sito citato in precedenza, i valori di accuratezza oscillano tra l'85% e il 95%, ma soltanto in riferimento ai corpora di inglese moderno[8], specificando, tra l'altro, che tali percentuali sono destinate a scendere di qualche punto, com'è ovvio, in relazione ai corpora di inglese antico; nessun dato, invece, viene dichiarato per i corpora relativi alle altre lingue.

Tutte queste problematiche condizionano l'efficacia della ricerca avanzata per parti del discorso, per categorie sintattiche ecc. Un ulteriore dato su cui riflettere: l'homepage di Google Ngram Viewer presenta di default un esempio di ricerca, in cui vengono messe a confronto le occorrenze presenti nei corpora in lingua inglese, relativamente al periodo 1800-2019, di Albert Einstein, Sherlock Holmes e Frankenstein[9]. Osservando la curva, si nota che le prime attestazioni nelle banche dati di Albert Einstein, per quanto oltremodo residuali (0,00000009%), sono databili 1816, frutto quindi di un errore o di una omonimia, dato che il fisico tedesco sarebbe nato oltre cinquant'anni dopo, nel 1879; e se qui la conferma del dato è semplice e intuibile, non lo è per quasi tutte le altre possibili ricerche, che indurrebbero l'utente a utilizzare dati che potrebbero essere, come in questo caso, inattendibili.

A differenza, però, di quanto visto in precedenza, le alternative a Google Ngram Viewer ci sono e sono facilmente reperibili in rete. Soltanto per l'italiano, banche dati, archivi, corpora differenziati tipologicamente, diafasicamente e diamesicamente, e ancora dizionari digitali storici e sincronici sono innumerevoli[10], ponendosi come ottimo ausilio per studiosi o appassionati che vogliano ricostruire la diffusione e l'uso di una parola, senza ricorrere, tuttavia, all'accattivante e quanto mai intellegibile grafico di Google Ngram Viewer. Si tratta di risorse che, però, condividono con Google analoghi problemi strutturali, ossia i modelli sono particolarmente funzionali per il trattamento dei testi contemporanei, molto meno per quelli di altre varietà, con particolare riferimento alle varietà storiche[11]; oppure, già nella fase di digitalizzazione, l'affidabilità degli strumenti per il riconoscimento ottico dei caratteri (OCR, *optical character recognition*) dipende in maniera determinante dalla qualità dell'inchiostro, dalle dimensioni, dalle fattezze e dal livello di

[8] I dati sono grossomodo in linea con i valori di accuratezza dei modelli sviluppati negli ultimi anni per il trattamento della lingua inglese, come testimonia il recentissimo contributo di Chiche e Yitagesu (2022).

[9] https://books.google.com/ngrams.

[10] Si vedano, a mo' d'esempio, gli *Scaffali digitali* presenti nel sito dell'Accademia della Crusca (https://accademiadellacrusca.it/), dove sono consultabili le maggiori risorse disponibili per l'italiano.

[11] Cfr. da ultimo Favaro et al. (2022).

usura della fonte cartacea[12]. La differenza, però, è che tali problemi vengono dichiarati e affrontati quotidianamente dalla comunità scientifica, con lo svantaggio che la mentalità accademica rallenta la divulgazione dei prodotti, fornendo così l'assist decisivo alle multinazionali private, che senza dubbio si pongono molti meno scrupoli.

2.3 Servizi per il machine learning: Google Colaboratory

Un altro strumento di Google, creato però per utenti già specializzati, è Google Colaboratory[13]. Si tratta di una piattaforma che permette di utilizzare un blocco note interattivo per eseguire codici scritti in Python, con la possibilità di condividerli immediatamente su Google Drive e GitHub (cfr. § 3) e di accedere alle *Graphic Processing Unit* (GPU) virtuali, processori che permettono di sostenere carichi di lavoro più impegnativi rispetto alle consuete *Central Processing Unit* (CPU), e quindi maggiormente adatti all'elaborazione di sistemi di machine learning che in genere utilizzano una ingente quantità di dati. Nell'ambito della linguistica computazionale, per esempio, gli strumenti allo stato dell'arte che permettono di costruire corpora annotati linguisticamente, come *UDPipe* (Straka e Straková 2017) e *Stanza* (Qi et al. 2020), progettati in riferimento allo standard internazionale delle *Universal Dependencies* (UD, De Marneffe et al. 2021), consigliano di addestrare i diversi modelli per l'annotazione utilizzando, appunto, le GPU, per evitare che il processo risulti laborioso o addirittura insostenibile per la macchina, a seconda dei dati da elaborare e dei parametri selezionati durante la configurazione iniziale del processo di addestramento.

In questi casi, l'impiego di Google Colaboratory può quindi risultare molto vantaggioso, perché consentirebbe di risolvere il problema senza dover acquistare e installare un hardware ad hoc. Vi sono, tuttavia, due osservazioni da fare: la prima riguarda il fatto che, come moltissimi altri servizi, la versione gratuita, accessibile a tutti, fornisce un numero limitato di GPU virtuali e di spazio in memoria, risolvibile soltanto pagando la versione Pro o persino Pro+; la seconda è che il blocco note virtuale usato per eseguire i codici è quello sviluppato dal progetto Jupyter[14], una organizzazione non profit che presta la risorsa anche ad Amazon SageMaker, un altro prodotto facente parte dei numerosi servizi "offerti" da Amazon Web Services[15]. Emerge dunque una realtà incontrovertibile: Google e Amazon, assieme a Microsoft, come vedremo nel prossimo paragrafo, creano costantemente profitto sfruttando risorse che nascono estranee alle logiche di mercato, ma che non hanno potuto (o voluto) contrastare lo strapotere economico dei colossi americani, divenendo soggetti alle loro dipendenze[16].

[12] Si vedano, ad esempio, i problemi nella digitalizzazione e nella strutturazione della versione online del *Grande Dizionario della Lingua Italiana* (Sassolini et al. 2021).
[13] https://colab.research.google.com/.
[14] https://jupyter.org/.
[15] https://aws.amazon.com/it/. Cfr. par. successivo.
[16] Cfr. § 3 per il caso analogo di GitHub.

2.4 Il cloud computing: Amazon Web Services e Microsoft Azure

Amazon Web Services (AWS) e Microsoft Azure[17] sono due delle più famose piattaforme di *cloud computing*, ovvero una serie di servizi, disponibili direttamente in rete, per il calcolo, l'archiviazione e analisi dei dati e altro. Come Google Colaboratory, AWS e Azure consentono di elaborare milioni e milioni di dati senza dover possedere centri di analisi o server fisici. Relativamente al machine learning, le piattaforme forniscono ambienti virtuali che l'utente può adoperare per costruire e addestrare i propri modelli. A differenza, però, di quanto osservato sopra per Google Colaboratory, il piano gratuito è disponibile soltanto per un periodo limitato dal momento della sottoscrizione, che rimane comunque obbligatoria, e i prezzi dei prodotti sono basati su una tariffa a consumo e, in particolare per quelli di Amazon, sull'accumulo: più ne usi, meno spendi.

I servizi di Amazon, infatti, sono organizzati a scompartimenti: soltanto per il machine learning sono sette, a loro volta organizzati in sottoservizi che sono correlati ad altri prodotti non facenti parte dello stesso pacchetto. Amazon Comprehend, per esempio, si occupa di elaborazione del linguaggio naturale (*natural language processing*, NLP) e permette il riconoscimento delle entità, l'analisi del sentiment e della sintassi e molte altre funzioni; per ottenere però la sintesi e per il riconoscimento della voce, risorse che fanno altrettanto parte dell'universo NLP, bisogna usare altri servizi (nello specifico, Amazon Polly e Amazon Transcribe). Questa organizzazione dei pacchetti non solo crea un enorme profitto per l'azienda di Seattle, ma anche una notevole dispersione, poiché ogni pacchetto ha un suo funzionamento specifico, comportando tra l'altro la necessità di competenze trasversali per poter sfruttare appieno le risorse. Tale ostacolo, però, sembrerebbe essere soltanto apparente: navigando sui siti delle due piattaforme, AWS e Azure, si ritrovano decine e decine di testimonianze sull'utilizzo quotidiano dei servizi, dalle grandi alle medie imprese, fino ad arrivare alle organizzazioni non profit. Da quelle stesse testimonianze, invece, l'ambito della ricerca scientifica sembrerebbe, almeno sulla carta, ancora capace di evitare o di impiegare con estrema parsimonia il *cloud computing* a pagamento[18].

3. Conclusioni

Nella sfera aziendale, sembrerebbe prevalere la logica del servizio virtuale su richiesta, malgrado i numerosi rischi e problemi dovuti all'uso di tali sistemi, che vanno dalla sicurezza, alla privacy, al divario digitale tra paesi ricchi e poveri, all'obsolescenza e alla migrazione dei dati, per citarne alcuni[19]. Invece, gli specialisti del machine learning, nell'ambito della ricerca, si affidano ancora, per la maggior parte, a risorse open-source, immagazzinando dati, sviluppando algoritmi, pacchetti e

[17] https://azure.microsoft.com/en-us/.

[18] Se è abbastanza semplice trovare in rete stime riguardanti l'utilizzo del cloud computing da parte delle aziende, non è altrettanto facile ricavare informazioni riguardanti l'utilizzo pubblico di tale risorsa. L'unico dato interessante e, se possibile, allarmante lo fornisce sempre AWS tramite il proprio sito: sarebbero già oltre 14000 gli istituti di pubblica istruzione, dalle scuole primarie alle università, che lo utilizzano o che organizzano corsi per insegnare agli studenti l'uso di tali strumenti. Non è possibile ricavare, tuttavia, alcun dato più specifico, in particolare sulla distribuzione areale dei suddetti istituti.

[19] Sul tema, si veda tra gli altri Dutta et al. (2013); più specificamente sui rischi legati alla ricerca, si veda Reddy et al. (2011).

strumenti per permettere a tutti gli utenti di creare risorse nuove o di aggiornare quelle esistenti, come avviene ad esempio per le banche dati della piattaforma UD (cfr. § 2.2), impiegate per l'addestramento dei modelli di annotazione, che sono state negli anni sviluppate da numerosi istituti di ricerca internazionali che si occupano di linguistica computazionale.

Anche in questo caso, però, c'è un problema alla fonte: gli sviluppatori usufruiscono anch'essi di supporti esterni per la condivisione delle risorse. Le banche dati sono infatti tutte scaricabili tramite GitHub, il servizio di hosting per eccellenza usato dagli sviluppatori di tutto il mondo per condividere i propri dati[20]; GitHub, però, sintetizza in sé un grande paradosso: nato come eccellenza open-source, è stato acquistato qualche anno fa da Microsoft, uno dei più grandi nemici del codice aperto[21].

GAFAM, si sa, penetra giorno dopo giorno, direttamente e indirettamente, spazi e ambiti non solo commerciali, generando di continuo profitto sotto la maschera del bene comune. Ciò, tra l'altro, ostacola fortemente lo sviluppo di alternative concrete e indipendenti ideate dal settore pubblico, che quando riesce a creare strumenti altrettanto validi e spesso migliori, rischia di dare alla luce prodotti inaccessibili agli utenti non specializzati, anche solo per il fatto che questi prodotti alternativi non vengono utilizzati in primo luogo dalle stesse istituzioni pubbliche che li realizzano: si pensi, per esempio, ai servizi per videoconferenze creati dal Consortium GAAR, associazione nata dalla collaborazione tra il Ministero dell'Istruzione, dell'Università e della Ricerca e varie importanti istituzioni come il CNR o l'INAF[22], che in tempi di pandemia, di didattica a distanza e di lavoro agile non sono mai riusciti a contrastare il monopolio di Google Meet, di Microsoft Teams o di Zoom. Il pericolo di affidarsi esclusivamente ai cinque colossi americani è dietro l'angolo; questo rischio è forse già realtà, se si pensa anche, oltre a quanto appena osservato, al costante utilizzo da parte delle università dei servizi di posta elettronica e per l'archiviazione dei dati, offerti in prima linea da Google.

Ma è giusto lasciare che GAFAM si insinui in tutto ciò che è relativo all'accademia e alle istituzioni pubbliche, dichiarare fin d'ora la resa incondizionata ai giganti del web e rinunciare all'autonomia della ricerca solo per l'immediato vantaggio dovuto all'abbattimento dei costi e lo sfruttamento di infrastrutture preesistenti? L'unica strada percorribile per sottrarre il monopolio a GAFAM è quella di imparare a rendere usufruibili e appetibili i prodotti sviluppati dall'università e dai centri di ricerca, anche al di fuori delle mura accademiche.

Bibliografia

Biber D., "Representativeness in corpus design", *Literary and Linguistic Computing*, 8, 1993, Oxford, Oxford University Press, pp. 243-257.

[20] https://github.com/.

[21] Rijtano R., "Microsoft compra GitHub, il 'social' degli sviluppatori", *La Repubblica*, 4 giugno 2018, https://www.repubblica.it/tecnologia/2018/06/04/news/microsoft_compra_github_il_social_degli_sviluppatori-198116594/.

[22] https://www.garr.it.

Chiche A., Yitagesu B., "Part of speech tagging: a systematic review of deep learning and machine learning approaches", *Journal of Big Data*, 9, 2022. Disponibile su: https://doi.org/10.1186/s40537-022-00561-y.

De Marneffe M. C., Manning C. D., Nivre J. and Zeman D., "Universal Dependencies", *Computational Linguistics*, 47(2), 2021, pp. 255-308.

Dutta, A., Peng, G.C. and Choudhary, A., "Risks in enterprise cloud computing: the perspective of IT experts", *Journal of Computer Information Systems*, 53 (4), 2017, pp. 39-48.

Favaro M., Biffi M., Montemagni S., "Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione", in Misuraca M., Scepi G., Spano M. (a cura di), *Proceedings of the 16th international conference on statistical analysis of textual data*, Vadistat press, Napoli, 2022, II vol., vol. I, pp. 392-399.

Klein L., "Hacking the Field: Teaching Digital Humanities with Off-the-Shelf Tools", *Transformations: The Journal of Inclusive Scholarship and Pedagogy*, 1, 2011, pp. 37-52.

Melvin J., Mike S., Quoc V. L., Maxim K., Yonghui W., Zhifeng C., Nikhil T., Fernanda V., Martin W., Greg C., Macduff H., Jeffrey D., "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation", *Transactions of the Association for Computational Linguistics*, 5, 2017, pp. 339–351.

Michaels, W. B., "The Beauty of a Social Problem (e.g. unemployment)", *Twentieth Century Literature*¸ 3-4, 2011, pp. 309-327.

Qi P., Zhang I., Zhang Y., Bolton J., Manning C. D., "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, July 5-10*, Association for Computational Linguistics, 2020, pp. 101-108.

Reddy K. V., Rao B. T., Reddy L. S. S., Kiran P. S., "Research Issues in Cloud Computing", *Global Journal of Computer Science and Technology*, 11, 2011. Disponibile online su: https://computerresearch.org/index.php/computer/article/view/794/794.

Sassolini, E., Biffi, M., De Blasi, F., Guadagnini, E., Montemagni, S., "La digitalizzazione del GDLI: un approccio linguistico per la corretta acquisizione del testo?", in Boschetti, F., Del Grosso A. M. & Salvatori E. (a cura di), *AIUCD 2021 - DHs for society: e-quality, participation, rights and values in the Digital Age. Book of extended abstracts of the 10th national conference*, Pisa, Associazione per l'Informatica Umanistica e la Cultura Digitale, pp. 159-166.

Stahlberg, F., "Neural Machine Translation: A Review", *Journal of Artificial Intelligence Research*¸ 69, 2020, pp. 343-418.

Straka M., Straková J., "Tokenizing, POS Tagging, Lemmatizing and Parsing UD2.0 with UDPipe", in *Proceedings of the CoNLL 2017: Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Association for Computational Linguistics, 2017, pp. 88-89.