La digitalizzazione del *GDLI*: un approccio linguistico per la corretta acquisizione del testo?

Eva Sassolini, Marco Biffi, Francesca De Blasi,

Elisa Guadagnini, Simonetta Montemagni

https://aiucd2021.labcd.unipi.it/wp-content/uploads/2021/01/a067.pdf

Il contesto della ricerca: il progetto TrAVaSI (**Tr**attamento **A**utomatico di **Va**rietà **S**toriche di **I**taliano)

TrAVaSI

- nasce dalla collaborazione tra l'Istituto di Linguistica Computazionale "A. Zampolli" del CNR e l'Accademia della Crusca
- è co-finanziato dalla Regione Toscana con le risorse del POR FSE 2014-2020 – Asse A Occupazione, all'interno di "GiovaniSì", il progetto regionale Toscano per l'autonomia dei giovani

Obiettivo

- costruire risorse e strumenti per il trattamento automatico di varietà storiche della lingua italiana, al momento pressoché inesistenti nel panorama nazionale e internazionale. L'obiettivo oggetto del presente contributo consiste nell'estrazione e strutturazione della conoscenza contenuta in dizionari storici digitali, al fine di creare i presupposti per funzionalità avanzate di navigazione e interrogazione

Il punto di partenza

- uno degli strumenti realizzati dall'Accademia della Crusca all'interno di altri progetti, ovvero la versione elettronica del *Grande Dizionario della lingua italiana* (GDLI), un prototipo della quale è attualmente consultabile negli *Scaffali digitali* del Sito Web dell'Accademia

La strutturazione e marcatura del GDLI nella sua versione informatizzata è oggetto del presente contributo

GDLI e www.gdli.it

- Il *GDLI* è un dizionario storico della lingua italiana, l'unico a coprire, per quanto con alcuni limiti, l'intero arco cronologico dalle origini fino a oggi. È pertanto uno strumento fondamentale per lo studio dell'italiano in diacronia.
- L'opera è suddivisa in 21 volumi, usciti tra il 1961 e il 2002, più 2 supplementi (usciti nel 2004 e nel 2008) e un volume di indici delle abbreviature delle fonti citate.
- L'attuale versione elettronica (www.gdli.it), resa disponibile in rete dall'Accademia della Crusca in seguito a un accordo con la casa editrice UTET, consente un'interrogazione *full text* per forme del testo ottenuto con applicazione di OCR senza nessuna collazione di controllo. L'approdo finale di una ricerca è il testo elettronico agganciato alla riproduzione in facsimile delle pagine.
- Vista la grande importanza dello strumento e le enormi potenzialità di ricerca seppure in una versione con errori di riconoscimento, l'Accademia della Crusca ha deciso di mettere a disposizione degli studiosi di lingua italiana anche questa versione provvisoria e ha avviato, di concerto con l'ILC, il progetto di implementazione che qui presentiamo, volto anche all'individuazione di procedure, metodi e strumenti più generalmente applicabili alla lessicografia elettronica, diretta (per la costruzione di dizionari) e indiretta (per la trasformazione di dizionari cartacei preesistenti).

GDLI: il dizionario originale

- Caratteristiche fisiche: il dizionario è stampato su carta non bianca e semilucida e conta 23 volumi di grande formato, per un totale di oltre 25.000 pagine
- Caratteristiche editoriali: ogni pagina del dizionario è divisa in tre colonne, il font è fitto e schiacciato, la dimensione carattere può arrivare fino a 6.5 punti, l'interlinea è molto piccola
- Caratteristiche lessicografiche: il testo è fortemente strutturato (ogni voce può avere da tre a otto campi), sono presenti diverse varietà di italiano, si riscontrano molte abbreviazioni e caratteri speciali

Il dizionario originale

- Il punto di partenza
- Il processo di strutturazione dei dati
- L’approccio linguistico
- Il supporto alla correzione manuale

DIVULGAZIONE — 889 — DIZIONE

2. Traduttore.
Montale, 3-112: Dopo aver consultato un suo calepino d'indirizzi s'era fatto indicare da un droghiere la casa da lui desiderata... Abitava dunque là Ponzio Macchi, il più instancabile e forse il più sottile dei suoi divulgatori stranieri?
3. Ant. Che celebra, che esalta.
Brandèlio, 2-24 (I-891): Voi séte divota e serva d'essa madama la reina e continova e chiara divolgatrice de le sue rare doti.
= Voce dotta, lat. tardo *divulgātor -ōris*, da *divulgāre* ‘divulgare’.

Divulgazióne (ant. *divolgazióne*), sf. Il divulgare, il rendere noto pubblicamente; diffusione di notizie, idee, dottrine, usanze, scritti.
Tasso, II-536: Io aspetto la ricompensa di quel dispiacere che mi ha portato la divolgazione de l'opere mie così maltrattate. *Davila*, 29: Io non posso indurmi ad affermarlo sopra la sola e molte volte fallace divulgazione della fama. *Tommaseo*, 1-138: Il Cioni mi manda la risposta che il Becchi fece alla sua lettera intorno all'edizione de' novellieri e de' comici. Ma io non vo' cooperare alla divulgazione di sozzure. *E. Cecchi*, 5-452: Mi ricordo certe figure che erano in voga, anni addietro, durante la prima e più entusiastica divulgazione del progresso scientifico ed industriale. *Brancati*, II-201: Ormai non aveva più fiducia che una sola parola, pronunciata a bassa voce in quel salotto, potesse sfuggire alla divulgazione.
– Ant. In senso concreto: notizia, voce, diceria.
Davila, 241: Ma molto di più di questa divulgazione, universalmente creduta falsa, empierono il Re di sospetto le lettere di Monsignore di San Goart.
2. In partic., con riferimento a certa produzione artistica o letteraria: esposizione facile, chiara e formulata in un linguaggio largamente comprensibile, di nozioni scientifiche e tecniche, di dottrine, di ideologie che divengono in tal modo accessibili a un pubblico assai più vasto della cerchia intellettuale che le ha elaborate (e può essere implicita, nel termine, una connotazione spregiativa, in relazione alla mancanza di originalità di tali opere e, soprattutto, a certa superficialità e approssimazione che a volte vi si riscontra).
E. Cecchi, 8-117: La letteratura di consumo corrente è francese: romanzi, libri gialli, periodici letterari e di varietà; collane filosofiche e psicologiche, con abbondanza di divulgazioni freudiane. *Serra*, I-271: Questo è infine il nostro carattere più vero, nella cultura, come nell'arte: la banalità, l'imprecisione; la grossolanità delle disposizioni generiche senza la vita e il rilievo dei particolari. È una specie di divulgazione, di livellamento democratico. *Gramsci*, 6-106: Perché non è nata una letteratura di divulgazione scientifica, come in Francia e negli altri paesi?
= Voce dotta, lat. tardo *divulgatiō -ōnis*, da *divulgāre* ‘divulgare’.

Divulghevolménte (*divolghevolménte*), avv. Ant. Pubblicamente.
Valerio Massimo volgar., I-89: Et allora udì che colui era la crudele morte di Cicilia e d'Italia, e che quando elli fosse sciolto di quelli legami sarebbe stragimento di molte cittadi; il quale sogno il dì seguente ridisse divolghevolmente.
= Comp. di un *divulghevole*, non attestato.

Divulsióne, sf. Letter. Il divellere, lo strappare.
Faldella, 3-470: Il giorno dopo la catastrofe, il paesaggio si fece veder limpido, come nulla di straordinario fosse accaduto. Le spighe dei campi, che pur sentivano sul collo le carezze del vento, e le foglie dei gelsi timorose della prossima divulsione non fecero neppure argomento del loro mormorio la tragedia del 2 giugno.
2. Medic. Dilatazione violenta, operata a mano o mediante appositi strumenti, di orifizi muniti di anelli sfinterici, che ne provoca la lacerazione rendendone così la contrazione impossibile (ed è praticata nella cura del cardiospasmo e delle ragadi anali).
Pansini, IV-202: ‘Divulsione’. In medicina significa dilatazione forzata... ‘Divulsione del piloro, del retto’.
= Voce dotta, lat. *divulsiō -ōnis* (da *divellĕre*: cfr. Divellere).

Divulso, v. Divelto[1].

Divulsóre, sm. Medic. Strumento chirurgico usato per praticare la divulsione.
= Voce dotta, deriv. da *divulso*, part. pass. di *divellere*: cfr. Divellere.

Dixitte, sm. Ant. Solo nella locuz. *Avere più anni del dixitte*: essere antichissimo.
Nomi, 8-14: Proverem se a resister sarà buona / la lor muraglia e quelle palafitte / che oggi mai han più anni del dixitte.
= Dal lat. *dixit*, 3ª pers. perf. ind. di *dīcere* ‘dire’ (pronunciato toscanamente *dìxitte*), frequente formula d'inizio del Salmo.

Dizaccherare (dial. *dinaccarare*), tr. (*dizàcchero*). Ant. Pulire un abito inzaccherato.
S. Bernardino da Siena, 858: Di verno infangasi e guastasi il vestimento da pici, ché s'involle nel fango come fa una porca, e poi vi perde uno dì a dizaccararlo.
= Comp. da *di-* con valore di separazione e *saccharo* (v.).

Dizeccolare e deriv., v. Dizzeccolare e deriv.

Diziferare (*dizziferare*, *dizifrare*), tr. (*dizìfero*). Ant. e dial. Decifrare.
G. C. Croce, 237: – Oh, questa è una cosa che la sanno tutti i nostri pecorari lassù. – È possibil questo? Io la tengo per una cosa molto intricata. – Orsù, io ve la voglio dizifferare or ora. *Chiari*, II-196: M'ero quasi adattata nell'animo mio alle dure mie circostanze, quando, a dizifrare questo altro enigma, m'arrivò da Napoli una lettera di D. Virginia.
= Variante dial. (di area settentr.) di *decifrare* (v.); cfr. *sifera* (*sifra*).

Diziferato (part. pass. di *diziferare*), agg. (*dizifrato*). Ant. e dial. Decifrato.
Chiari, II-92: Ecco dizifrato l'enigma che il conte Fristano nella sua lettera indicato m'avea.

Dizigòtico, agg. (plur. m. *-ci*). Biol. Derivante dalla fecondazione di una cellula-uovo diversa da quella da cui ha origine il gemello; dicoriale.
= Voce dotta, comp. da *di-* (dal gr. δι- ‘doppio’) e *zigote* (v.).

Dizionarièsco, agg. (plur. m. *-chi*) Raro. Pedante, libresco.
Bacchelli, 2-XXII-253: In Italia, non sono mai occorse quelle revisioni e importazioni di gusto e fondazioni di nuovi canoni, che l'indole accademica e dizionariesca della letteratura francese e l'indole troppo fresca e malleabilissima della tedesca hanno periodicamente prodotto.
= Deriv. da *dizionario*, sul modello di *libresco*.

Dizionàrio, sm. Opera che raccoglie, accompagnate da una definizione, le parole di una lingua, nella loro totalità sincronica e diacronica o limitatamente a particolari età della tradizione linguistica, o a particolari autori, o a particolari categorie di parole (neologismi, barbarismi, sinonimi, tecnicismi delle varie scienze, arti, mestieri): l'ordine in cui le parole sono disposte è generalmente alfabetico, più di rado etimologico; non infrequente è l'ordine metodico (per cui le parole sono raggruppate secondo la loro affinità concettuale). – *Dizionario bilingue*: quello in cui i vocaboli e le locuzioni di una lingua sono definiti con quelli corrispondenti di un'altra lingua (ed è strumento essenziale per le traduzioni). – *Dizionario storico, geografico, biografico, degli autori e delle opere*: raccolta ordinata di nomi propri di personaggi e luoghi e di titoli di opere, accompagnati dalle notizie relative. – *Dizionario enciclopedico*: riunisce i caratteri sia del dizionario linguistico (in quanto tiene conto anche del patrimonio lessicale di una lingua), sia del dizionario storico, geografico, scientifico, tecnico.
P. F. Giambullari, 5-234: È egli però vero che tutte queste voci siano aramee? Certissimo, gli rispos[e] io. E che se ne mostra? mi soggiunse egli. Ed io: I dizionari stessi caldei ed ebrei che si trovano oggi stampati. *Fiamma*, 296: Ha scritto ancora gli Avvertimenti morali sopra tutta la Bibia; dato buon principio al Dizionario teologico. *Serdonati*, 10-125: Egli primo di tutti aveva composto la grammatica della lingua giapponese e copiosi dizionari. *Magalotti*, 1-28: In oggi... ogni mestiero ha la sua Poliantea e il suo gran dizionario da sé. *Milizia*, II-123: Fece un corso compito d'architettura. Fu il primo a fare un dizionario di quest'arte. *Da Ponte*, XXIII-41: Era quel libro un dizionario, tedesco e italiano: a' lochi indicati lessi queste tre parole: « Ich liebe Sie »; e trovai che significavano « Io amo voi ». *Tramater* [s. v.]: Il titolo di ‘Vocabolario’ o ‘vocabolista’ non si applica propriamente che a' puri ‘dizionarii’ di parole. ‘Dizionario’ è d'un significato più esteso, poichè comprende in generale non solo i ‘dizionarii’ delle lingue, ma anche i ‘dizionarii’ storici, e quelli delle scienze e delle arti. *Tommaseo*, 3-I-141: Nel dizionario ch'io desidererei, gioverebbe distinguere i modi schietti del popolo da quelli degli scrittori, che sono sovente come di lingua morta. *De Sanctis*, I-77: Eccoti un vascello. In un dizionario per categorie, alla voce vascello troverai infinite voci toscane per nominare le sue parti ed i suoi attrezzi. *Jovine*, 5-174: Stava consultando un pesante dizionario e ne aveva appoggiato il dorso al listello inferiore della libreria.
– Figur. Scherz.
Monti, 3-6-103: Trimuglio... sapeva intero a mente / delle buone creanze il dizionario. *Guerrazzi*, 5-98: Andando innanzi di questo passo tu devi essere arrivato presto a comporre il dizionario della lingua degli occhi.
2. Disus. Lessico; linguaggio.
Vico, 251: Idee uniformi nate appo intieri popoli tra essoloro non conosciute debbon avere un motivo comune di vero... Ond'esce il dizionario mentale, da dar l'origini a tutte le lingue articolate diverse, col quale sta concepita la storia ideal eterna che ne dia le storie in tempo di tutte le nazioni. *G. Gozzi*, 1-113: Le prime parole che noi tutti impariamo... ci vengono dagli orecchi nella memoria col mezzo delle balie, delle cameriere... A poco a poco ci nasce opportunità di vedere i congiunti e gli amici della famiglia, e di qua si accresce il nostro dizionario nella mente di vocaboli più colti e più gentili. *Alfieri*, V-2-853: Il dizionario politico vuol essere rinnovato del tutto: tutti i nomi più sacrosanti essendo ora stati contaminati da tante bocche e penne servili, bisogna alla libertà, alla proprietà, ai diritti dell'uomo, alle leggi, ad ogni cosa insomma dar nuovi nomi. *De Sanctis*, 11-154: L'amore ha anche nella poesia la sua storia: prima passionato e rozzo, a poco a poco si fa gentile... e poi va sfumando nel manierato, nel galante, nel convenzionale, e la freschezza del primo linguaggio diviene un dizionario di moda. *Dossi*, 521: Né il mio dizionario or rifiuta, come altre volte, tutte quelle vivaci espressioni che danno il colore ed i muscoli ad un discorso.
3. Dimin. *Dizionariétto*.
Carducci, III-15-116: Appunto per le belle fu dato quel dizionarietto mitologico, che dalla seconda edizione in poi accompagnò le canzonette.
– Spreg. *Dizionariàccio*.
Mazzini, I-704: Le vostre dichiarazioni sul principio e le lunghe mie due lettere d'istruzioni, tornatemi insieme ad un dizionaruccio del quale non vi giovaste mai, ...promettevano, stipulavano altro.
= Voce dotta, lat. mediev. *dictiōnārium* (da *dictiō -ōnis* ‘modo di dire’ sul modello di *vocabularium* ‘vocabolario’). *Vocabolario* è sinonimo di *dizionario linguistico* (ma non è sinonimo di *dizionario enciclopedico*).

Dizionarista, sm. e f. (plur. m. *-i*). Compilatore di un dizionario.
Roberti, VI-2-151: Era d'uopo in primo luogo provare..., ciò che non fa il dizionarista, che il lusso veramente arricchisca colui che fa lusso.

Dizióne[1], sf. Il dire, discorso (anche riferito a singole e concrete espressioni linguistiche, orali o scritte).
Speroni, 1-4-111: Questa terza condizione del nostro verso, cioè la rima, essendo diligente inquisizione della moltitudine ed identità delle lettere che entrano in essa sillaba sopra la quale casca la rima, par che sia condizione propria della dizione in quanto umana; cioè in quanto terminata e significante. *Monti*, IV-262: Io ho stimato bene di non tenermi che quelle [varianti] che inducono variazione nella sentenza o che migliorano la dizione. *Carducci*, III-18-15: L'infranciosamento, a cui la patria di Dante lasciossi andare con gli ultimi Medici e si disfrenò co' Lorenesi, corrose l'incarnato e la forza natia della dizione. *Baldini*, I-720: Per quanto impegnato in una dizione largamente panneggiata, non gli sentii cadere in tutto il discorso sillaba in fallo, non ci fu lo strappo d'una sola congiunzione.
– Gramm. Ant. *Parti della dizione*: parti del discorso.
Varchi, 8-2-111: Aristotile, dove favella verso il fine delle parti della dizione e divide il nome in più specie, non intende di quel propio di che al presente favelliamo noi.
2. L'aspetto formale di un'opera letteraria, quello cioè che concerne la scelta e disposizione delle parole. – Per estens.: stile.
Speroni, 1-4-158: Il divino Alighieri... [chiama] stil tragico ogni poetica dizione, la qual senza imitare senta e tegna del grande. *Abati*, 168: Persio può anche imitarsi in qualche tratto di magnifica dizione e di giudizioso insultamento; ma non deve nella secca maniera del suo fraseggiare e nella erudizione astrusa costituirne esempio. *Muratori*, 5-I-121: Nella sentenza poi, o vogliam dire ne' sentimenti e nella dizione, o sia nelle frasi e parole con cui descrivono i poeti le cose, infinitamente ancora si perfeziona la natura. *Foscolo*, XI-2-316: Raramente avviene che chi scrive moltissimo possa alla fretta aggiungere eleganza di dizione, vigore d'idee condensate e calore di stile. *Leopardi*, II-32: La lingua francese è poverissima. Quindi la necessità di metafore, di metonimie, di catacresi, di mille figure di dizione. *De Sanctis*, 7-162: Questo difetto artistico, questa mancanza di senso artistico, si riflette in ciò che si usa dir forma, nella dizione, nella lingua. *Gramsci*, 6-364: L'azione drammatica è tutta in questa tenuità, ...è condotta con candore, con dizione semplice e quasi scialba.
– Raro. Sintassi.
Leopardi, I-302: Così pure accade nel latino, che i più antichi (scrittori) sono i più facili, e di dizione più somigliante di gran lunga alla greca.
3. Lettura ad alta voce o recitazione di testi letterari. – Anche (con speciale riferimento ad attori, conferenzieri, annunciatori radiofonici e televisivi): arte di leggere o recitare un testo in modo da farne percepire interamente a chi ascolta la bellezza poetica o l'efficacia persuasiva (con esattezza di pronuncia, giusta intonazione di voce, espressività).
Oriani, X-19-85: Anch'io ho sognato ieri sera nel piccolo teatro della mia città, ascoltando Masini forse per l'ultima volta... La sua dizione, il suo accento, la sua voce ripetono l'antico incanto. *Ojetti*, II-435: Vorrebbe che questi giovani, di popolo i più, studiassero anche storia, letteratura, dizione, trucco, per anni, prima di cimentarsi alla ribalta. *E. Cecchi*, 1-171: La sua dizione era la più brillante e tuttavia la più modesta, retta da un senso delle pause, dei silenzi, da un rispetto dell'arabesco scenico, che si intendevano meglio sapendo che cotesta attrice proveniva dalla danza.
– In partic.: modo di articolare i suoni, pronuncia.
C. E. Gadda, 7-129: Noi udiamo invece il toscano serbare con amico animo e cogliere agevolmente le sdrucciole, tesoreggiarle nella glottide, liberarle serenamente all'aere quando opportunità ne richieda: opportunità..., cioè un finissimo senso musicale, nella dizione spontanea.

Una pagina del GDLI

Divulgazióne (ant. *divolgazióne*), sf. Il divulgare, il rendere noto pubblicamente; diffusione di notizie, idee, dottrine, usanze, scritti.
Tasso, II-536: Io aspetto la ricompensa di quel dispiacere che mi ha portato la divolgazione de l'opere mie così maltrattate. *Davila*, 29: Io non posso indurmi ad affermarlo sopra la sola e molte volte fallace divulgazione della fama. *Tommaseo*, 1-138: Il Cioni mi manda la risposta che il Becchi fece alla sua lettera intorno all'edizione de' novellieri e de' comici. Ma io non vo' cooperare alla divulgazione di sozzure. *E. Cecchi*, 5-452: Mi ricordo certe figure che erano in voga, anni addietro, durante la prima e più entusiastica divulgazione del progresso scientifico ed industriale. *Brancati*, II-201: Ormai non aveva più fiducia che una sola parola, pronunciata a bassa voce in quel salotto, potesse sfuggire alla divulgazione.
– Ant. In senso concreto: notizia, voce, diceria.
Davila, 241: Ma molto di più di questa divulgazione, universalmente creduta falsa, empierono il Re di sospetto le lettere di Monsignore di San Goart.
2. In partic., con riferimento a certa produzione artistica o letteraria: esposizione facile, chiara e formulata in un linguaggio largamente comprensibile, di nozioni scientifiche e tecniche, di dottrine, di ideologie che divengono in tal modo accessibili a un pubblico assai più vasto della cerchia intellettuale che le ha elaborate (e può essere implicita, nel termine, una connotazione spregiativa, in relazione alla mancanza di originalità di tali opere e, soprattutto, a certa superficialità e approssimazione che a volte vi si riscontra).
E. Cecchi, 8-117: La letteratura di consumo corrente è francese: romanzi, libri gialli, periodici letterari e di varietà; collane filosofiche e psicologiche, con abbondanza di divulgazioni freudiane. *Serra*, I-271: Questo è infine il nostro carattere più vero, nella cultura, come nell'arte: la banalità, l'imprecisione; la grossolanità delle disposizioni generiche senza la vita e il rilievo dei particolari. È una specie di divulgazione, di livellamento democratico. *Gramsci*, 6-106: Perché non è nata una letteratura di divulgazione scientifica, come in Francia e negli altri paesi?
= Voce dotta, lat. tardo *divulgatiō -ōnis*, da *divulgāre* ‘divulgare’.

Una voce del GDLI

L’output del sistema di OCR: il punto di partenza

- Il testo restituito dal sistema di OCR è in formato non standard (word) e di difficile gestione
- Principali problematiche incontrate:
  - sul piano della formattazione: es. mancata individuazione dell’entrata; mancato riconoscimento dei paragrafi di definizione e/o di esempi; mancato riconoscimento dei paragrafi di commento etimologico; cattiva segmentazione delle voci, ecc.
  - sul piano testuale:

1) cattiva segmentazione	1.1) introduzione di spazi bianchi dove non ci sono
	1.2) univerbazione di parole separate
2) grafie scorrette	2.1) omissione di singoli caratteri o sequenze
	2.2) cattivo riconoscimento di singoli caratteri o di sequenze di caratteri
	2.3) cattivo riconoscimento di caratteri che produce l’inserimento di caratteri aggiuntivi

Tabella 1: tipologie di errori

L’output del sistema di OCR: il punto di partenza
Alcuni esempi

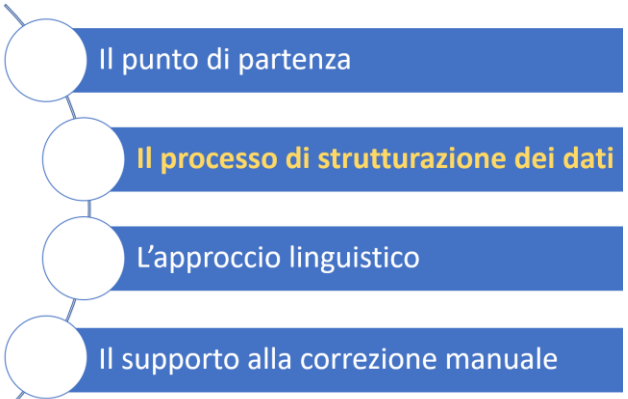
N.	Originale cartaceo	Testo OCR
1	**Amminoazobenzène** (*aminoazobenzène*), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di *giallo d’anilina*: cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).	Am mi no a z ob e nz è ne (*aminoazobenzène*), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di *giallo d’anilina:* cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).
2	**Assolare¹**, tr. (*assólo*). Disus. Rendere solo. – *Assolare una carta*: tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da *solo* (v.). **Assolare²**, tr. (*assólo*). Esporre al sole; rendere soleggiato. = Deriv. da *sole* (v.). **Assolare³** (*assuolare*), tr. (*assòlo* o *assuòlo*). Disporre a strati. = Deriv. da *suolo* (v.).	**Assolare¹**, tr. *[assòlo).* Disus. Rendere solo. - *Assolare una carta*: tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da *solo* (v.). **Assolare²**, tr. *[assòlo).* Esporre al sole; rendere soleggiato. = Deriv. da *sole* (v.). **Assolare³** *[assuolare),* tr. *[assòlo o assuòlo).* Disporre a strati. = Deriv. da *suolo* (v.).
3	**Ammacchiare¹**, rifl. (*m’ammàcchio, t’ammacchi*). Raro. Nascondersi nella macchia. *B. Davanzati*, I-136: Floro s’ammacchiò: vedendo poi presi i passi dell’uscita, s’uccise.	Ammacchiate¹) rifl. *{m’ammàcchio, Vammacchi).* Raro. Nascondersi nella macchia. *B. Davanzali,* I-136: Floro s’ammacchiò: vedendo poi presi i passi dell’uscita, s’uccise.
4	**Attendista,** agg. e sm. e f. (plur. m. *-i*). Neol. Chi evita di prendere posizione (e resta in attesa degli avvenimenti, riservandosi di decidere secondo il loro svolgersi). = Fr. *attentiste* (1941), da *attendre* ‘attendere’. **Attenditóre,** agg. e sm. (femm. *-trice*). Ant. Che attende, aspetta.	=Deriv.·da·*attendere*.¶ Attendista, ·agg.· e· sm.· e· f.· (plur.· m.· *-i*).· Neol. Chi· evita· di· prendere· posizione· (e· resta· in· attesa· degli· avvenimenti,· riservandosi· di· decidere secondo·il·loro·svolgersi).¶ =·Fr.·*attentiste*·(1941),·da·*attendre*·‘attendere’.·¶ Attenditóre, agg.·e·sm.·(femm.·*-trice*).·Ant.·¶ Che·attende,·aspetta.¶

Problematiche sul piano del testo

Problematiche sul piano della formattazione

Metodi e tecnologie utilizzate per il processo di strutturazione: i dati di input

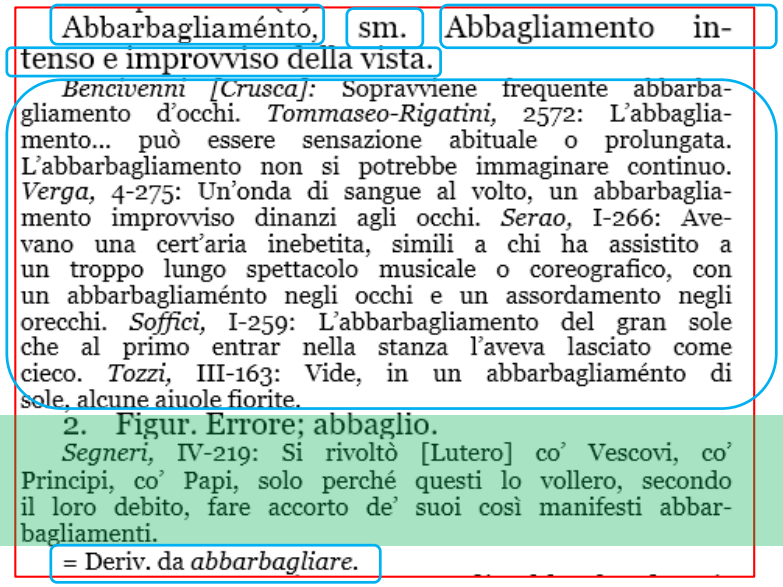
- L'output del sistema di OCR
  - la fase di acquisizione non è parte del nostro progetto
    - Impossibilità di impedire il ripetersi delle casistiche di errori più ricorrenti, individuando possibili strategie di miglioramento, cosa fattibile potendo avere un feedback dalla procedura di OCR
- La presenza di un formato dei dati digitali non standard e la presenza di errori distribuiti di vario tipo ha condizionato la scelta dell'approccio all'estrazione



Metodi e tecnologie utilizzate per il processo di strutturazione dei dati

Il processo di estrazione automatica

- L'analisi dei dati ha reso necessaria una strategia di estrazione a regole personalizzate con livelli di raffinamento progressivi:
  - In una prima fase la segmentazione del testo in porzioni corrispondenti al corpo dell'intera entrata
  - successivamente si procede con il riconoscimento degli altri campi dell'entrata grazie all'identificazione delle caratteristiche distintive, tradotte in vincoli di corretta attribuzione
  - Il processo è impostato in modo incrementale



Intera voce

Singoli campi

Altro senso con citazione interna

Metodi e tecnologie utilizzate per il processo di strutturazione dei dati: l'approccio linguistico

È possibile impiegare «intelligenza linguistica» per mettere a punto una strategia di trattamento efficace degli errori, che cerchi di ridurre al minimo la necessità di interventi manuali?

- Modularità nell'approccio:
  - Separare le strategie di individuazione degli errori da quelle di correzione automatica
  - Diversificare le strategie che interessano l'intero testo da quelle mirate al singolo campo
- Ad oggi abbiamo sviluppato procedure d'intervento sugli errori a tutto testo e nel campo lemma

Metodi e tecnologie utilizzate per il processo di strutturazione dei dati: l'approccio linguistico

Errori a tutto testo:

- Per avere un quadro generale di questi errori, abbiamo creato una «sezione *gold*»: una selezione casuale di 30 colonne, che sono state corrette manualmente
- L'analisi delle risultanze di questa selezione rappresenta lo studio preliminare su chi poggiano le strategie di correzione adottate

LIVELLO DI CORREZIONE	AREA DI TESTO INTERESSATA DALL'ERRORE	DESCRIZIONE DELL'ERRORE	STRATEGIA DI INDIVIDUAZIONE/CORREZIONE DELL'ERRORE
livello 0	singolo carattere	presenza di un carattere non alfabetico (o non appartenente all'alfabeto italiano)	• analisi degli errori della "sezione *gold*" (sostituzioni 1:1)
livello 1	sequenza di 2 o più caratteri (ogni carattere di per sé è ammissibile)	sequenza di caratteri non ammessa in lingua italiana	• analisi degli errori della "sezione *gold*" • "criterio fonotattico"
livello 2	parola (caratteri e sequenze di caratteri di per sé sono ammissibili)	parola errata	• analisi degli errori della "sezione *gold*" • confronto con altri *corpora*

Metodi e tecnologie utilizzate per il processo di strutturazione dei dati: l'approccio linguistico

Strategia di individuazione/correzione dell'errore: le risultanze della sezione *gold*

OCR errato	Lezione corretta
*	'
*	'
*	'
*	'
*	'
*	'
*	'

Livello 0
Es.* > '

OCR errato	Lezione corretta
??	zz
??	zz
??	zz
??	zz
??	zz
??	zz
??	zz
??	zz
??	zz

Livello 1
Es. ?? > zz

OCR errato	Lezione corretta	Passo errato	Passo corretto	Campo della voce	Fonte	vol_pa
pass,	pass.	part. pass,	part. pass.	definizione	Gold	01_20_1
pass,	pass.	part. pass,	part. pass.	definizione	Gold	01_20_1
pass,	pass.	part. pass,	part. pass.	definizione	Gold	03_488_1
pass,	pass.	part. pass,	part. pass.	definizione	Gold	06_339_3
		part. pass,	part. pass.	definizione	Gold	08_263_1
		part. pass,	part. pass.	definizione	Gold	16_1042_2

Livello 2
Es. pass, > pass.

Metodi e tecnologie utilizzate per il processo di strutturazione dei dati: l'approccio linguistico

Strategie di individuazione di nessi consonantici certamente errati: criterio fonotattico

Es. impcrvietà

Nessi biconsonantici non ammessi in italiano

bf bg bh bm bp bq bt bv bz cb cd cg cp cv db df dh dl dp dq dt dv dz fb fd fh fm fp fq fs fv fz gc gf gp gq gs gt gv gz hb hc hd hf
ht hv hz lh md mf mg mh mr ms mt mv mz nb nh nm np pb pc pd pf pg ph pm pq pv qb qc qd qf qg qh ql qm qn qp qr qs qt qv qz
vb vc vd vf vg vh vm vn vp vq vs vt vz zb zd zf zg zh zl zm zn zp zq zr zs zv

Nessi triconsonantici non ammessi in italiano

bbb bbc bbd bbf bbg bbh bbm bbn bbp bbq bbs bbt bbv bbz bcb bcc bcd bcf bcg bch bcl bcm bcn bcp bcq bcr bcs bct bcv bcz bdb bdc bdd
bdq bdr bds bdt bdv bdz bfb bfc bfd bff bfg bfh bfl bfm bfn bfp bfq bfr bfs bft bfv bfz bgb bgc bgd bgf bgg bgh bgl bgm bgn bgp bgq bg
bhf bhg bhh bhl bhm bhn bhp bhq bhr bhs bht bhv bhz blb blc bld blf blg blh bll blm bln blp blq blr bls blt blv blz bmb bmc bmd bmf
bmq bmr bms bmt bmv bmz bnb bnc bnd bnf bng bnh bnl bnm bnn bnp bnq bnr bns bnt bnv bnz bpb bpc bpd bpf bpg bph bpl bpm bp
bqb bqc bqd bqf bqg bqh bql bqm bqn bqp bqq bqr bqs bqt bqv bqz brb brc brd brf brg brh brl brm brn brp brq brr brs brt brv brz bsb b
bsp bsq bsr bss bst bsv bsz btb btc btd btf btg bth btl btm btn btp btq btr bts btt btv btz bvb bvc bvd bvf bvg bvh bvl bvm bvn bvp bvq b
bzf bzg bzh bzl bzm bzn bzp bzq bzr bzs bzt bzv bzz cbb cbc cbd cbf cbg cbh cbl cbm cbn cbp cbq cbr cbs cbt cbv cbz ccb ccc ccd ccf c
ccz cdb cdc cdd cdf cdg cdh cdl cdm cdn cdp cdq cdr cds cdt cdv cdz cfb cfc cfd cff cfg cfh cfl cfm cfn cfp cfq cfr cfs cft cfv cfz cgb cg
cgp cgq cgr cgs cgt cgv cgz chb chc chd chf chg chh chl chm chn chp chq chr chs cht chv chz clb clc cld clf clg clh cll clm cln clp clq c
cmf cmg cmh cml cmm cmn cmp cmq cmr cms cmt cmv cmz cnb cnc cnd cnf cng cnh cnl cnm cnn cnp cnq cnr cns cnt cnv cnz cpb cp
cpp cpq cpr cps cpt cpv cpz cqb cqc cqd cqf cqg cqh cql cqm cqn cqp cqq cqr cqs cqt cqv cqz crb crc crd crf crg crh crl crm crn crp crq cr
csg csh csl csm csn csp csq csr css cst csv csz ctb ctc ctd ctf ctg cth ctl ctm ctn ctp ctq ctr cts ctt ctv ctz cvb cvc cvd cvf cvg cvh cvl cv
cvz czb czc czd czf czg czh czl czm czn czp czq czr czs czt czv czz dbb dbc dbd dbf dbg dbh dbl dbm dbn dbp dbq dbr dbs dbt dbv dbz dc
dcn dcp dcq dcr dcs dct dcv dcz ddb ddc ddd ddf ddg ddh ddl ddm ddn ddp ddq dds ddt ddv ddz dfb dfc dfd dff dfg dfh dfl dfm dfn dfp
dgd dgf dgg dgh dgl dgm dgn dgp dgq dgr dgs dgt dgv dgz dhb dhc dhd dhf dhg dhh dhl dhm dhn dhp dhq dhr dhs dht dhv dhz dlb dlc
dlq dlr dls dlt dlv dlz dmb dmc dmd dmf dmg dmh dml dmm dmn dmp dmq dmr dms dmt dmv dmz dnb dnc dnd dnf dng dnh dnl dnm dnr
dpb dpc dpd dpf dpg dph dpl dpm dpn dpp dpq dpr dps dpt dpv dpz dqb dqc dqd dqf dqg dqh dql dqm dqn dqp dqq dqr dqs dqt dqv dqz
drn drp drq drr drs drt drv drz dsb dsc dsd dsf dsg dsh dsl dsm dsn dsp dsq dsr dss dst dsv dsz dtb dtc dtd dtf dtg dth dtl dtm dtn dtp dtq
dvf dvg dvh dvl dvm dvn dvp dvq dvr dvs dvt dvv dvz dzb dzc dzd dzf dzg dzh dzl dzm dzn dzp dzq dzr dzs dzt dzv dzz fbb fbc fbd fbf

- Il punto di partenza
- Il processo di strutturazione dei dati
- L'approccio linguistico
- Il supporto alla correzione manuale

Metodi e tecnologie utilizzate per il processo di strutturazione dei dati: l'approccio linguistico

Strategia di individuazione/correzione dell'errore nel campo lemma

- controllo delle «testatine»:
  - confronto del lemma con l'intervallo alfabetico segnalato dalle "testatine" (vale a dire le entrate della prima voce della prima colonna e dell'ultima voce della terza colonna, segnalati nell'intestazione di ogni pagina del *GDLI*): grazie a questo confronto è possibile formulare delle ipotesi di correzione della parte iniziale del lemma – è così possibile per esempio correggere l'errato "bfdicotomia" in "broncotomia".
- confronto con le entrate del *Dizionario della lingua italiana* «Tommaseo-Bellini»

Metodi e tecnologie utilizzate per il processo di strutturazione dei dati: l'approccio linguistico

Le strategie descritte hanno mostrato la necessità di un approccio modulare

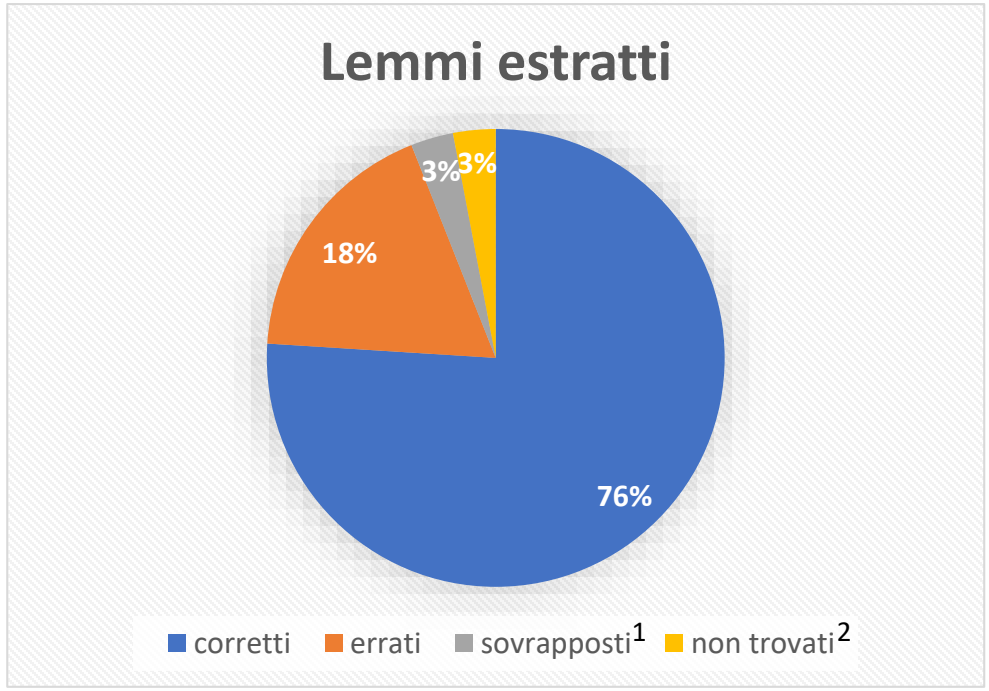
- Gli interventi correttivi si collocano in fasi e tempi diversi del processo di estrazione:
  - L'output del sistema di OCR per gli errori generalizzati (a tutto testo)
  - i testi dei campi della voce in fase di estrazione, valutando la combinazione di diverse caratteristiche
- Anche le procedure software diventano modulari e utilizzano diverse risorse linguistiche nelle fasi mirate di estrazione di singoli campi/porzioni di campo

Realizzazione di strumenti e di risorse linguistiche a supporto della correzione manuale: valutazione dei risultati

- Una parte degli errori descritti è stata corretta automaticamente ma resta necessaria una revisione manuale post-processing
- Implementazione di strategie e metodi di supporto alla correzione e sistema di revisione e riallineamento successivo dei dati estratti
- Parallelamente alla fase di estrazione dei dati sono stati implementati meccanismi di annotazione puntuale:
  - creazione di un file di report che segnala la mancanza di campi obbligatori o se il loro ordine non è rispettato, con ubicazione precisa del punto;
  - quando è possibile impostare un'indagine più puntuale del dato, i file di report sono finalizzati al controllo delle soluzioni già inserite in fase di parsing di estrazione, così da facilitare il lavoro di revisione manuale
- Impostazione di strumenti software per supportare la correzione assistita del lemmario:
  - sulla base delle evidenze emerse nello studio della sezione *gold* e dei criteri di individuazione degli errori sopra descritti è stato stilato un preciso protocollo di correzione automatica e manuale che si avvale di un opportuno strumento software di supporto dedicato

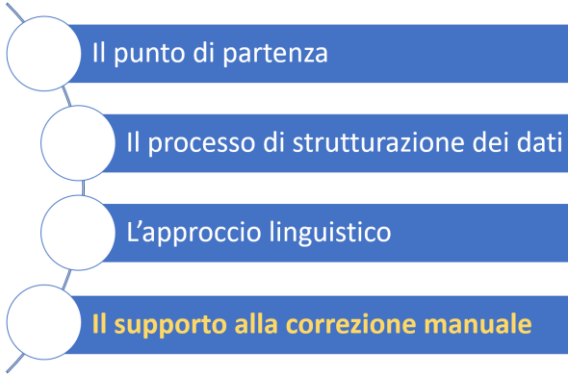
Realizzazione di strumenti e di risorse linguistiche di supporto alla correzione manuale

- Tutti i file di report sono organizzati e gestiti da un software che assiste la correzione manuale
- Il testing dello strumento software di supporto è stato fatto impostando una prima sessione di lavoro:
  - sul primo volume
  - relativamente alla correzione del lemma
- Nell'immagine accanto è riportata la sintesi



1 ***sovrapposti***: conseguenza dell'errata interpretazione dello strumento morfologico del sistema di OCR, es. «aggranchito» per «aggronchito»

2 ***non trovati***: con struttura dell'entrata diversa dalla norma, es. manca la categoria grammaticale anche nell'originale



Conclusioni

- Il lavoro già oggi rappresenta un valido prototipo di intervento in situazioni in cui la mole e la qualità dei dati da elaborare rendono necessaria la messa a punto di procedure *ad hoc*
- Il primo scopo del lavoro è portare l'output del sistema di OCR in un formato standard di rappresentazione, che consenta la strutturazione del vocabolario e quindi moltiplichi le sue possibilità di consultazione e interoperabilità
- In prospettiva è quindi pensabile utilizzare altri strumenti software di analisi per affinare ulteriormente la marcatura dei dati