Annotazione semantica e visualizzazione di un corpus di corrispondenze di guerra

Beatrice Dal Bo, Francesca Frontini, Giancarlo Luxardo
Praxiling UMR 5267
Université Paul-Valéry Montpellier 3 - CNRS
France
name.surname@univ-montp3.fr

Abstract

English. This paper introduces Corpus 14, a corpus of correspondences between French soldiers and their relatives during the Great War. We describe the digital edition and its encoding in TEI, as well as the ongoing activities related to indexing and referencing of places, persons and other named entities, undertaken in order to represent the network of correspondences by means of a geovisualisation.

Italiano. Questo articolo presenta Corpus 14, un corpus di corrispondenze tra i soldati francesi della Grande Guerra e le loro famiglie. Descriviamo l'edizione digitale e la sua codifica in TEI, nonché i lavori attualmente in corso per indicizzare e referenziare luoghi, persone ed altre entità nominate, al fine di poter rappresentare la rete di scambi epistolari attraverso una visualizzazione grafica di tipo spaziale.

1 Introduzione

Il progetto Corpus 14, iniziato in concomitanza con il centenario della Grande Guerra, nasce dalla volontà di studiare la lingua delle persone comuni all'inizio del XX secolo, ed in particolare degli scriventi *peu lettrés*, che potremmo tradurre, seguendo la letteratura italiana, con *semicolti* (D'Achille, 1994). In questo contesto si tratta dei *Poilus*, soldati francesi della Grande Guerra, spesso provenienti dalle campagne e da contesti rurali, ancora in parte dialettofoni, che si confrontano spesso per la prima volta con il testo scritto. Se lo studio delle competenze linguistiche e pragmalinguistiche è alla base della raccolta delle loro corrispondenze, tali documenti si dimostrano essere una fonte interessante anche per altre discipline, con interessanti informazioni di carattere storico, geografico e culturale. In particolare l'interesse si è focalizzato su due ambiti:

- lo studio della Grande Guerra e della sua eredità in termini di memoria sociale, e delle trasformazioni da essa prodotte,
- l'evoluzione degli usi linguistici, in particolare per quanto riguarda l'influenza delle varietà regionali (in particolare per le zone come il Sud della Francia o la Bretagna, caratterizzate da diglossia), o lo sviluppo di un socioletto comune, il cosiddetto *argot des poilus*.

Utilizzando materiale proveniente da archivi pubblici, nonché documenti donati da eredi al progetto, Corpus 14 si compone ad oggi (versione 2.0[1]) di 37 scriventi, provenienti da 11 regioni diverse, per un totale di 1.797 lettere e circa 500.000 parole. I criteri di selezione del corpus sono stati i seguenti:

- la selezione di scriventi che non hanno completato la formazione elementare,
- la preferenza per le corrispondenze complete, o che per lo meno permettessero di seguire gli scriventi su un arco temporale lungo, e che potessero dunque dare luogo a reti di corrispondenze complesse. Al momento Corpus 14 è costituito di 11 reti di corrispondenze, raggruppate per zona geografica e nominate secondo i luoghi di origine (si veda Figura 1)

[1] https://hdl.handle.net/11403/corpus14

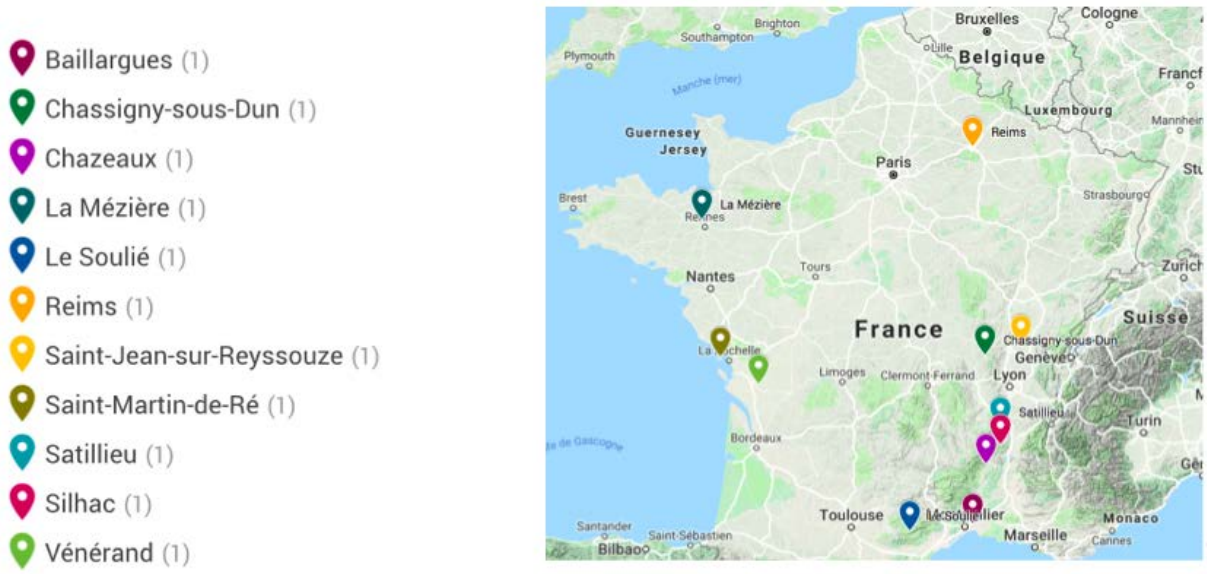

Figura 1: Localizzazione dei luoghi di origine dei soldati di Corpus 14.

Tali criteri di selezione fanno dei fondi di Corpus 14 una collezione unica nel suo genere. Tuttavia la sua realizzazione si ispira anche a progetti faro nella comunità delle edizioni digitali di corrispondenze, molti dei quali dedicati agli epistolari di personaggi illustri[2], con poche eccezioni, come: "Digitising experiences of migration: the development of interconnected letter collections" di Moreton e Nesi, 2013-2014[3]. Inoltre, per la tematica il progetto può essere accostato ad altri omologhi sviluppati in diversi paesi europei in occasione del centenario della Prima Guerra Mondiale, come l'italiano "Voci della Grande Guerra"[4] ed il britannico "Letters from the First World War"[5].

2 L'edizione digitale

L'edizione digitale si è avvalsa di pratiche già ben stabilite, come la trascrizione diplomatica del testo, l'allineamento tra i facsimile delle cartoline o delle lettere e la loro codifica precisa (con precisazioni sulla leggibilità del testo).

Per quanto riguarda la codifica, si è fatto appello allo standard della Text Encoding Initiative (TEI[6]). In particolare le trascrizioni sono state effettuate in modo da permettere la descrizione della struttura logica del testo, nonché delle caratteristiche di leggibilità del supporto fisico. Per ogni lettera sono state realizzate due versioni (fedele e normalizzata all'ortografia corrente).

L'applicazione di questo schema di annotazione XML alla tipologia testuale in oggetto è stato facilitato dall'esistenza di un gruppo di lavoro sulle corrispondenze in seno alla comunità TEI[7]. In particolare si è fatto ricorso agli elementi *TEIheader*, *correspDesc* e *CorrespAction* (introdotti nella versione 2.8.0 delle specifiche TEI P5).

Per quanto riguarda la distribuzione, Corpus 14 è reso disponibile in diverse modalità di accesso che garantiscono la fruizione da parte di tipologie di utenti diverse. Da una parte si è voluto fornire un'interfaccia di esplorazione[8] ed analisi del testo attraverso la piattaforma di testometria TXM (si

[2]Uno dei progetti più noti in questo senso è *Mapping the Republic of Letters*, http://republicofletters.stanford. edu/; per una ricognizione più completa di tali progetti si veda (Stadler et al., 2016)

[3]http://lettersofmigration.blogspot.com; per ulteriori informazioni si veda (Moreton et al., 2014; Moreton, 2016)

[4]http://www.vocidellagrandeguerra.it/

[5]https://www.nationalarchives.gov.uk/education/resources/letters-first-world-war-1915/

[6]https://www.tei-c.org

[7]*Special Interest Group della TEI* sulle corrispondenze https://tei-c.org/activities/sig/correspondence/

[8]http://textometrie.univ-montp3.fr/

veda la Figura 2)[9]. Allo stesso tempo i sorgenti TEI sono scaricabili dalla piattaforma Ortolang[10], che garantisce l'interoperabilità dei dati, la loro preservazione e la loro reperibilità (tramite il protocollo OAI-PMH).

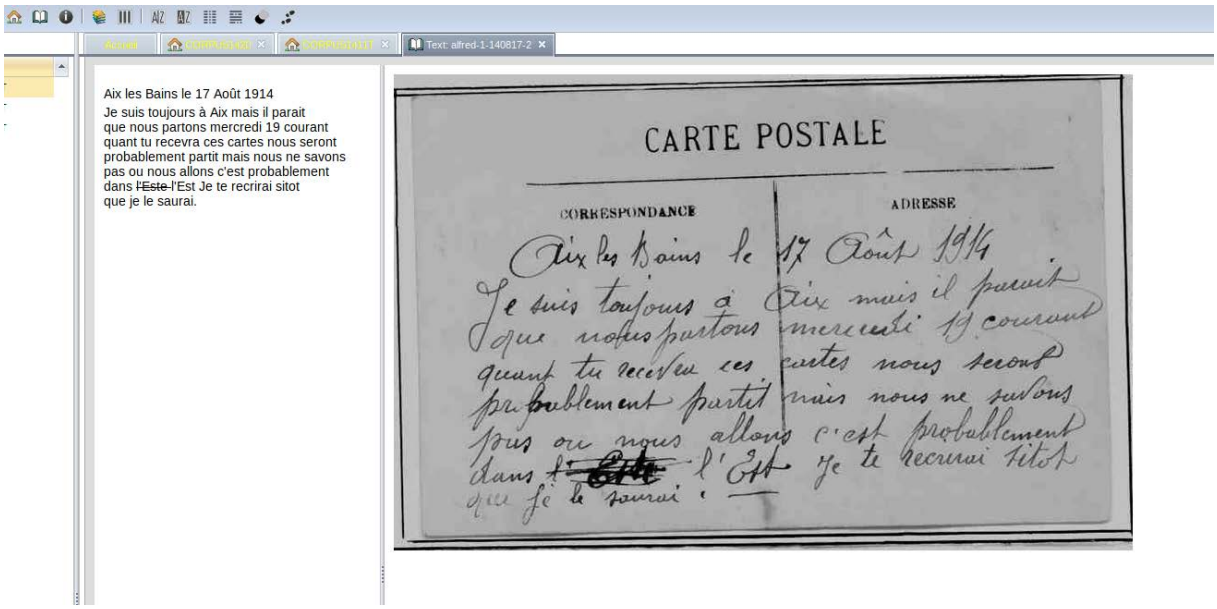

Figura 2: L'interfaccia di esplorazione del corpus TXM.

3 L'indicizzazione semantica dei testi

Una volta realizzata la prima versione dell'edizione digitale, si è posto il problema di arricchire e indicizzare i testi, ed in particolare di creare indici a persone, luoghi, organizzazioni citate. Tali indici, collegati ai riferimenti nel testo, dovranno poi essere arricchiti e collegati con l'informazione corrispondente disponibile.

L'indicizzazione dei testi, che per ora esiste solo su due reti di corrispondenze (*Chazeaux* e *Le Soulié*), è stata condotta secondo le buone pratiche della codifica in TEI, che si sono delineate anche nel contesto di gruppi di lavoro francesi facenti riferimento al consorzio CAHIER[11], come il progetto *Testaments de poilus*[12].

In particolare le menzioni di luoghi, persone e organizzazioni sono state dapprima annotate nel testo di ogni lettera (sia nei metadati della corrispondenza che nel corpo della lettera), utilizzando gli elementi TEI *persName*, *placeName*, *orgName*. Si è inoltre scelto di annotare oltre ai nomi propri anche stringhe di testo aventi nel contesto della lettera dei referenti univoci, usando l'elemento *rs*. L'annotazione è stata effettuata in maniera ricorsiva, dunque un'espressione come "les cousins de Cicignan" è stata annotata come una *rs*, contenente un *placeName*.

In seguito ogni menzione è stata referenziata con l'attributo *ref* e un codice univoco. Tale codice rinvia a tre indici, file separati contenenti delle liste di persone, luoghi, organizzazioni (*listPerson*, *listPlace*, *listOrg*). Per il referenziamento a DBpedia si è utilizzato il sistema di riconoscimento automatico di entità nominate REDEN Online (Résolution et Désambiguïsation d'Entités Nommées) (Frontini et al., 2016), con postcorrezione manuale.

Infine, tali liste sono state dove possibile arricchite con informazioni addizionali in nostro possesso (come le date e i luoghi di nascita e morte delle persone, scriventi o solo menzionate, il loro grado di

[9]TXM è uno strumento per l'esplorazione e l'analisi statistica di corpora testuali, sviluppato dall'ENS di Lione. Permette tra le altre cose l'import di testi annotati in TEI. Si veda http://textometrie.ens-lyon.fr.

[10]ORTOLANG, Outils et Ressources pour un Traitement Optimisé de la LANGue è la piattaforma francese per la pubblicazione delle risorse linguistiche, ora integrata all'infrastruttura CLARIN ERIC. https://www.ortolang.fr/

[11]CAHIER, Corpus d'Autueur pour les Humanités Numériques, è un consorzio di progetti aliati all'infrastruttura Huma-Num, che si occupa di edizioni digitali principalmente in TEI. Si veda https://cahier.hypotheses.org/

[12]https://testaments-de-poilus.huma-num.fr/

parentela, ecc.). Per quanto riguarda i luoghi si è fatto ricorso alla georeferenziazione e all'aggiunta di link al database geografico esterno GeoNames, oltre a quanto già referenziato su DBpedia. In alcuni casi, toponimi non presenti nelle basi sono stati individuati e localizzati. In alcuni casi, toponimi non presenti nelle basi sono stati individuati e localizzati.

4 Visualizzazione

Attualmente in corso è lo sviluppo di una piattaforma di visualizzazione, che permetterà di esplorare le corrispondenze in maniera geolocalizzata[13]. Come si può vedere dalla Figura 3, l'interfaccia permette di selezionare gli scambi epistolari di una stessa rete familiare per data, proiettando sulla carta ad esempio la lettera di un soldato e la risposta della moglie. Nella visualizzazione i segnaposto indicano il luogo di invio della lettera, mentre le bandierine indicano i luoghi citati nella lettera. La visualizzazione è realizzata in modo da sfruttare al massimo lo standard TEI recuperando i *placeName* con interrogazioni basate su XQuery (sia all'interno dei metadati corrispDesc che nel corpo della lettera) e utilizzando la geocodifica degli indici. In questo modo, una volta terminata, la piattaforma potrà essere riutilizzata come base per altri progetti con lo stesso formato di annotazione[14].

Figura 3: L'interfaccia di visualizzazione e geolocalizzazione delle corrispondenze.

5 Analisi

Numerose analisi sono state condotte su Corpus 14, si cita in particolare il volume collettivo curato da Agnès Steuckardt (Steuckardt, 2015a), nel quale sono analizzati vari aspetti linguistici di queste corrispondenze, fra cui la punteggiatura (Steuckardt, 2015b), l'ortografia (Pellat, 2015), il lessico (Luxardo, 2015) e la lingua regionale (Géa, 2015). Ricordiamo inoltre altri studi riguardanti aspetti morfosintattici (Steuckardt and Dal Bo, 2018) o discorsivi (Dal Bo and Wionet, 2018). Una tesi di dottorato è attualmente in corso di completamento (Dal Bo, 2019).

[13]La rappresentazione delle reti di corrispondenze attraverso visualizzazioni è una componente tipica di questo tipo di progetti. Si vedano ad esempio (O'Leary and Moreton, 2017); *Visual Correspondence*, http://www.correspondence.ie; *Mapping the Republic of Letters*, http://republicofletters.stanford.edu/; *Early Modern Letters Online*, http://emlo.bodleian.ox.ac.uk/home; *Clavius on the Web*, http://claviusontheweb.it/.

[14]Il progetto di interfaccia è stato realizzato da studenti del corso di laurea specialistica in informatica dell'Università di Genova, sotto la supervisione della prof. Marina Ribaudo.

Per quanto riguarda gli aspetti spaziali, l'analisi dei luoghi citati nelle corrispondenze dei soldati ha permesso di mettere in evidenza il fatto che questi evochino nelle lettere in maniera prevalente luoghi legati alla loro vita familiare, alla casa e agli affetti, e molto meno luoghi legati alla guerra e al fronte (come già messo in evidenza da (Dal Bo and Wionet, 2018; Gibelli, 2016)). La proiezione su una carta geografica delle informazioni geografiche e temporali delle corrispondenze permette inoltre di seguire gli spostamenti dei soldati al fronte e delle donne rimaste all'interno del Paese. Gli spostamenti di quest'ultime, più raramente studiati, potranno essere inoltre paragonati a quelli delle donne appartenenti a classi sociali superiori durante lo stesso periodo storico.

Bibliografia

Paolo D'Achille. 1994. L'italiano dei semicolti. In *Storia Della Lingua Italiana*, Einaudi, Torino, volume 2, pages 41–79.

Beatrice Dal Bo. 2019. *Aux Frontières de La Norme : Usages Linguistiques de Scripteurs Peu Lettrés Dans Des Correspondances de La Grande Guerre*. Ph.D. thesis, Université Paul-Valéry Montpellier 3.

Beatrice Dal Bo and Chantal Wionet. 2018. Alleviare l'assenza : La modalità ingiuntiva in alcune lettere di donne peu-lettrées durante la Grande Guerra. In Fabio Caffarena and Nancy Murzilli, editors, *In Guerra Con Le Parole. Il Primo Conflitto Mondiale Dalle Testimonianze Scritte Alla Memoria Multimediale*, Fondazione Museo Storico del Trentino, Trento, pages 187–201.

Francesca Frontini, Carmen Brando, and Jean Gabriel Ganascia. 2016. REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pages 193–197.

Jean-Michel Géa. 2015. Le dialecte dans l'écriture de la Guerre : la parte absente ? In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 53–65.

Antonio Gibelli. 2016. *La guerra grande: Storie di gente comune*. Laterza, Bari.

Giancarlo Luxardo. 2015. Fréquences des colis et marmites : comment mesurer la languitude ? In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 113–123.

Emma Moreton, Niall O'Leary, and Patrick O'Sullivan. 2014. Visualising the Emigrant Letter. *Revue européenne des migrations internationales* 30(vol. 30 - n°3 et 4):49–69. https://doi.org/10.4000/remi.7081

Emma Louise Moreton. 2016. *The Emigrant Letter Digitised: Markup and Analysis*. D_ph, University of Birmingham.

Niall O'Leary and Emma Moreton. 2017. The Migrant Letter Digitised: Visualising Metadata. *Journal of Cultural Analytics* https://doi.org/10.22148/16.013

Jean-Christophe Pellat. 2015. Les graphies des Poilus, loin des canons orthographiques. In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 67–77.

Peter Stadler, Marcel Illetschko, and Sabine Seifert. 2016. Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>. Journal of the Text Encoding Initiative (Issue 9). https://doi.org/10.4000/jtei.1433

Agnès Steuckardt, editor. 2015a. *Entre village et tranchées: l'écriture de poilus ordinaires*. Inclinaison, Uzès.

Agnès Steuckardt. 2015b. Sans point ni virgule. In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 91–100.

Agnès Steuckardt and Beatrice Dal Bo. 2018. Avoir été ou être allé ? Évolution d'une concurrence, d'après des corpus lettrés et peu lettré. In Peter Blumenthal and Denis Vigier, editors, *Études Diachroniques Du Français et Perspectives Sociétales*, Peter Lang, Berlin, page 295.