Strumenti digitali e collaborativi per le Scienze dell'Antichità
a cura di Paolo Mastandrea

WordNets per lingue classiche

Riccardo del Gratta
(Istituto di Linguistica Computazionale "A. Zampolli", CNR-ILC, Pisa, Italia)

Abstract The WordNet for Ancient Greek (AGWN) is presented and illustrated both as a stand-alone semantic net and as a participant in a more complex net of wordnets for historical and modern languages. Two applications of AGWN carried out within the Memorata Poetis project are described: in the first application, the AGWN is used to investigate multilingual synonyms; in the second one, AGWN is used to extract Greek synonyms in order to classify epigrams in terms of similar content.

Keywords WordNet. Classical Languages. Ancient Greek WordNet. Perseus. NLP.

1 Introduzione

Il presente contributo descrive brevemente il processo di creazione di una rete di *WordNets* per le lingue classiche e ne illustra due diverse destinazioni d'uso, entrambe incentrate sul progetto *Memorata Poetis*.[1]

Nella prima destinazione d'uso, partendo da testi poetici greci manualmente annotati a livello di temi e motivi si espandono le correlazioni tra parole nel testo e temi/motivi con cui i testi sono stati annotati aggiungendo sinonimi e relazioni semantiche sia alle prime che ai secondi. La seconda destinazione applica la rete semantica a epigrammi greci. Si cerca di raggruppare epigrammi che contengono termini che siano tra loro sinonimi o in relazione ipo/iperonimica. Entrambe le applicazioni sfruttano il modello concettuale delle reti semantiche: queste forniscono termini tra loro correlati a un insieme di testi da analizzare validando o meno le annotazioni o i raggruppamenti effettuati manualmente.

1 *Memoria Poetica e Poesia della Memoria,* progetto PRIN 2010/2011.

Antichistica 14
DOI 10.14277/6969-182-9/ANT-14-9 | Submission 2017-08-14
ISBN [ebook] 978-88-6969-182-9 | ISBN [print] 978-88-6969-183-6

2 WordNet e *WordNets*

WordNet (Fellbaum 1998) è una risorsa lessico-semantica il cui scopo principale è quello di formalizzare l'apparato teorico per creare e gestire una rete di concetti espressi da parole.

Per gli scopi del presente contributo, è sufficiente descrivere solo i blocchi principali del modello, rimandando ancora a Fellbaum (1998) per una completa ed esaustiva descrizione sia del modello completo che delle molteplici applicazioni di WordNet in diversi campi della Linguistica Computazionale.

Una tipica serializzazione del modello concettuale di WordNet è la trasformazione degli oggetti del modello in tabelle di un database relazionale. Le quattro tabelle principali e le relazioni che le connettono sono schematicamente mostrate nella fig. 1.

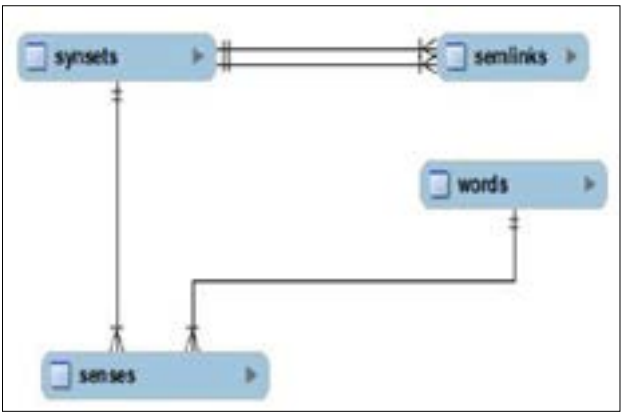

Figura 1. Le tabelle principali del modello WordNet (serializzato in un database)

La tabella 'words' contiene i singoli lemmi, come ad esempio *arm*, *limb, build up*, ecc. mentre i diversi sensi che i singoli lemmi possono assumere sono elencati nella tabella 'senses'. Questa tabella infatti contiene, ad esempio, il senso di *arm* sia come 'arto umano' (human limb) che come 'prepararsi per un'azione militare'. Per chiarezza, occorre dire che la tabella 'senses' non contiene le glosse dei sensi, i.e. 'arto umano' e 'prepararsi per un'azione militare', ma solo gli identificativi di queste glosse che, per contro, sono gestite dalla tabella dei 'synsets'.

La tabella 'synsets' è il cuore del modello. Oltre ad aggiungere le glosse ai sensi dei lemmi, essa svolge un doppio ruolo fondamentale: identifica univocamente i sensi aggiungendovi la parte del discorso (PoS) e raggruppa i sensi (corredati di PoS) in insiemi di sinonimi (il nome synset deriva dalla contrazione di *synonym sets*). Il senso del lemma *arm*, come verbo, inteso come 'prepararsi per un'azione militare' è sinonimo di *build up*, *fortify* e *gird*.

Infine la tabella 'semlinks' contiene le relazioni semantiche tra synsets: *arm* come arto umano è un iponimo di *limb*, inteso come una delle quattro appendici del corpo umano. La fig. 2 riporta graficamente gli esempi sopra descritti.

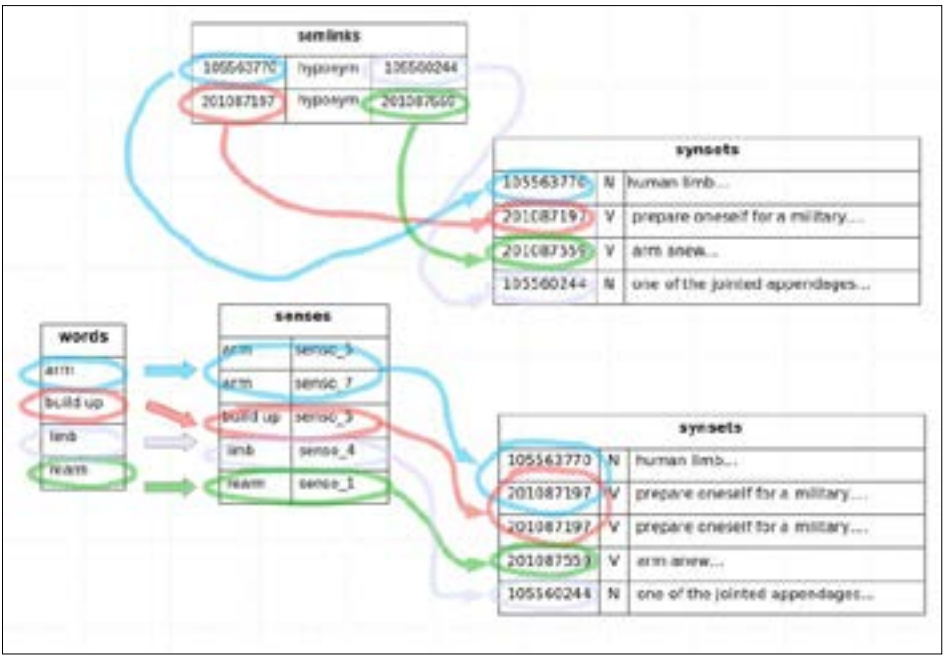

Figura 2. Parole, sensi, synsets e relazioni semantiche

Sebbene WordNet sia stato pensato espressamente per concettualizzare la lingua inglese (in letteratura viene citato come Princeton WordNet), il modello è generico e può essere applicato a lingue diverse, sia europee (Vossen 1998) che semitiche (Rodriguez et al. 2008). Seguendo il lavoro in Minozzi (2009), Bizzoni et al. (2014) hanno codificato la creazione di un database lessico-semantico per il greco antico basato sul modello WordNet.

3 Ancient Greek WordNet (AGWN)

Come mostrato in Bizzoni et al. (2014), la costruzione iniziale della WordNet per il greco antico si basa su dizionari bilingui digitalizzati greco-inglese forniti dal progetto Perseus (`http://www.perseus.tufts.edu`). La metodologia è quella di estrarre coppie di parole greco-inglese e confrontare la parola inglese con la lista di lemmi contenuti in Princeton WordNet. Se la

parola inglese è trovata, i suoi synsets vengono assegnati alla corrispondente parola greca.

Essendo il processo di creazione un processo di bootstrapping, ovvero di estrazione non controllata di dati, il risultato risulta alquanto impreciso: la polisemia indotta dall'inglese porta, ad esempio, la parola greca ἔχω a essere mappata su oltre 170 parole inglesi, tra cui *cut*, *make*, *break*; mentre se la parola inglese della coppia non è presente su Princeton WordNet, la parola greca viene esclusa, fatto, questo, che ha come conseguenza di limitare la copertura della WordNet per il greco antico al 30% del lessico totale.

Al momento attuale, la WordNet del greco antico è mappata su latino, italiano e inglese, in modo da avere una rete di wordnet interconnesse; anche questa mappatura, come quelle in Vossen (1998), Rodriguez et al. (2008), sfrutta l'inglese come lingua pivot, ovvero come lingua scelta per connettere tra di loro le altre lingue, per cui, di conseguenza, il greco e il latino o l'italiano sono interconnessi attraverso l'equivalente senso inglese.

4 Applicazione di AGWN a testi poetici greci annotati con temi/motivi

Lo scopo di questa applicazione è verificare da un punto di vista semantico se l'annotazione di un dato testo poetico greco con un particolare tema/motivo (in latino) sia suffragata da un'evidenza semantica.

Nell'applicazione si associano statisticamente una o più parole greche del testo con il tema/motivo con cui il testo è stato manualmente annotato. Questa lista di associazioni (le quali sono coppie bilingue greco-latine) viene controllata su AGWN per capire se la parola greca e quella latina sono mappate sullo stesso synset inglese. In altre parole, se la parola greca e quella latina sono sinonimi multiligui.[2] Il fatto che le parole greche e latine siano mappate sullo stesso synset inglese significa che sono entrambe traducenti dello stesso concetto ed in questo caso l'annotazione è molto significativa. La metodologia è ampiamente descritta in Boschetti (2014), a cui si rimanda anche per ulteriori casi interessanti. Di seguito è riportato un caso prototipico di sinonimia multilingue.

Da un testo si estrae la coppia termine greco/tema (in latino) {γραῦς, anus} che risulta mappata sullo stesso synset inglese che esprime il concetto di "donna vecchia".[3] In questo caso l'annotazione del testo con il

2 Spesso la parola greca e quella latina non sono sinonimi multilingui, ma c'è comunque evidenza di una connessione a livello di iperonimi: la parola greca è sinonima multilingue di una parola latina coiponima di quella sotto indagine, oppure gli iperonimi della coppia sono sinonimi multilingui.

3 In Princeton WordNet 3.0 il synset è identificato dal numero univoco 110377021.

tema/motivo *Anus* risulta potenziata dal fatto che la parola greca statisticamente più correlata al tema/motivo (γραῦς) esprime esattamente lo stesso concetto espresso dalla parola latina *anus*.

In aggiunta, la lista di sinonimi greci e latini che si estraggono dal synset comune permettono una associazione tra una qualsiasi parola greca sinonima di γραῦς e una qualsiasi parola latina sinonima di *anus:*

{γραίδιον, γραῖα, γραῦς, τηθία} → {*anicla*, *anicula*, *anicilla*, *anucella*, *anus*}

5 Applicazione di AGWN a testi epigrafici greci

La seconda applicazione d'uso è il raggruppamento di epigrammi greci. Attraverso AGWN, l'applicazione permette di raggruppare epigrammi che contengono termini tra loro sinonimi o in relazione (ipo)iperonimica tra di loro. Ad esempio i tre epigrammi, AG 9,492,493,494, sotto

- Κεῖτο δ' ὁμοῦ σάκος, ἔγχος, <ἄορ>, θώρηξ, κόρυς, ἵππος.
- Ἀσπίς, τόξα, βέλεμνα, κόρυς, ξίφος, ἄλκιμον ἔγχος.
- Ἰός, τόξα, σάκος, κυνέη, δόρυ, φάσγανα, θώρηξ.

trattano argomenti similari e possono essere raggruppati come ricorrenza lessico-semantica nella versificazione epigrafica che è un aspetto centrale del progetto Memorata Poetis.

Infatti AGWN rileva automaticamente la sinonimia tra σάκος e ἀσπίς, nel significato di scudo (shield); ξίφος e φάσγανα, intesi come spada (blade, sword) e ξίφος e ἄορ ancora come spada (blade, sword).

6 Conclusioni

In conclusione, si è cercato di illustrare come la WordNet per il greco antico, intesa sia come rete semantica a sé stante, sia come rete interconnessa a reti semantiche per le lingue classiche e moderne, sia in grado di aiutare l'annotazione e la classificazione di testi in greco antico. L'annotazione di testi poetici a livello di temi e motivi risulta potenziata quando le parole greche contenute nel testo e maggiormente correlate alla parola latina esprimente il tema/motivo sono semanticamente connesse; gli epigrammi risultano più facilmente classificabili per temi trattati grazie all'estrazione di sinonimi in essi contenuti. Nel primo caso la WordNet per il greco antico coopera con le altre reti semantiche, specificatamente quella latina e quella inglese, mentre nel secondo caso la rete è usata come risorsa lessico-semantica a sé stante.

Bibliografia

Bizzoni, Yuri et al. (2014). «The Making of Ancient Greek WordNet». *Proceedings of the 9th Conference on Language Resources and Evaluation* (Reykjavik, 26th-31st May 2014). Paris: ELRA, 1140-7.

Boschetti, Federico; Del Gratta, Riccardo (2014). «Computer Assisted Annotation of Themes and Motifs in Ancient Greek Epigrams. First Steps». Basili Roberto; Lenci, Alessandro; Magnini, Bernardo (eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014* (Pisa, 9-10 December 2014). Pisa: Pisa University Press, 83-6.

Fellbaum, Christiane (1998). *WordNet. An Electronic Lexical Database.* Cambridge (MA): The MIT Press.

Minozzi, Stefano (2009). «The Latin WordNet Project». Anreiter, Peter; Kienpointner, Manfred (eds.), *Latin Linguistics Today = Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik* (Innsbruck, 4-9 April 2009). Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck Bereich Sprachwissenschaft, 137: 707-16.

Rodríguez, Horacio et al. (2008). «Arabic WordNet. Current State and Future Extensions». *Proceedings of the Fourth International Global WordNet Conference. GWC 2008* (Szeged, Hungary, 22nd-25th January 2008). Amsterdam: IOS Press, 387-406.

Vossen, Piek (1998). *EuroWordNet. A Multilingual Database with Lexical Semantic Networks.* Norwell (MA): Kluwer Academic Publishers.