

LL(O)D and NLP Perspectives on Semantic Change for Humanities Research

Florentina Armaselu [a,*], Elena-Simona Apostol [b], Anas Fahad Khan [c], Chaya Liebeskind [d], Barbara McGillivray [e,f], Ciprian-Octavian Truică [g], Andrius Utka [h], Giedrė Valūnaitė Oleškevičienė [i] Marieke van Erp [j]

[a] *Luxembourg Centre for Contemporary and Digital History (C²DH), University or Luxembourg, Luxembourg*
E-mail: florentina.armaselu@uni.lu
[b] *Computer Science and Engineering Department, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Romania*
E-mail: elena.apostol@upb.ro
[c] *Istituto di Linguistica Computazionale «A. Zampolli», Consiglio Nazionale delle Ricerche, Italy*
E-mail: fahad.khan@ilc.cnr.it
[d] *Department of Computer Science, Jerusalem College of Technology, Jerusalem, Israel*
E-mail: liebchaya@gmail.com
[e] *Theoretical and Applied Linguistics, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, United Kingdom*
E-mail: bm517@cam.ac.uk
[f] *The Alan Turing Institute, United Kingdom*
E-mail: bmcgillivray@turing.ac.uk
[g] *Computer Science and Engineering Department, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Romania*
E-mail: ciprian.truica@upb.ro
[h] *Centre of Computational Linguistics, Vytautas Magnus University, Kaunas, Lithuania*
E-mail: andrius.utka@vdu.lt
[i] *Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania*
E-mail: gvalunaite@mruni.eu
[j] *DHLab, KNAW Humanities Cluster, Amsterdam, Netherlands*
E-mail: marieke.van.erp@dh.huc.knaw.nl

Author contributions: F.A., sections 1, 2, 3, 5, 6, 8; E.S.A., section 5; A.F.K., section 4; C.L., section 5; B.M., section 5; C.O.T., section 5; A.T., section 5; G.V.O., section 3; M.V.E., section 7. All the authors critically revised and approved the final version of the manuscript submitted to the Journal.

Abstract. The paper presents an overview of the LL(O)D and NLP methods, tools and data for detecting and representing semantic change, with main application in humanities research. Its aim is to provide the starting points for the construction of a workflow and set of multilingual diachronic ontologies within the humanities use case of the COST Action *Nexus Linguarum, European network for Web-centred linguistic data science*, CA18209. The survey focuses on the essential aspects needed to understand the current trends and to build applications in this area of study.

Keywords: linguistic linked open data, natural language processing, semantic change, ontologies, humanities

1. Introduction

The detection of semantic change in diachronic corpora and representing how a concept has changed over time as linked data is a core challenge on the intersection of digital humanities (DH) and Semantic Web (SW). Semantic Web technologies have already been used successfully in humanistic initiatives such as the Mapping the Manuscripts project [1] and in Pelagios [2]. They facilitate the creation, publication and interlinking of FAIR (Findable, Accessible, Interoperable and Reusable) datasets. In particular, the use of a common data model, common formalisms and common vocabularies in linked data helps to render datasets more interoperable; the use of readily available technologies such as the query language SPARQL also helps to make such data more (re-)usable. Semantic change data can be highly heterogeneous and potentially include linguistic, historic, bibliographic and geographical information, and linked data, with its shared data model, is well suited to handling this. For instance, the lexical aspect of such data is already served by the existing OntoLex-Lemon vocabulary and its extensions, and there also exist numerous vocabularies and datasets dealing with bibliographic metadata, historical time periods and geographic locations. In addition, the Web Ontology Language (OWL) and associated reasoning tools allow for basic ontological reasoning to be carried out on such data (this is useful for dealing with different classes of things referred to by word senses). Although important advances in the development of natural language processing (NLP) methods and tools for extracting historical entities and modelling diachronic linked data, as well as in the field of Linguistic Linked (Open) Data (LL(O)D)[1], have been made so far [3–5], there is a need for a systematic overview of this growing area of investigation. A few literature surveys and overview papers on the state of the art in lexical semantic change detection in NLP exist (e.g. [6–9]), but none addresses the intersection of this line of research with LL(O)D research. In particular, previous work has generally tended to focus on how to detect semantic change (in both corpora, e.g., [10], and linked data ontologies, e.g., [11]) but hasn't looked in depth at how to model and publish semantic change datasets in Linked Open Data (LOD) that result, at least in part, from these detection methods[2].

The contribution of this paper is a literature survey intended to consider these areas together. We posit that to better contextualise and target the combination of NLP and LL(O)D techniques for detecting and representing semantic change, the main workflow implied in the process should be taken into account. The term *semantic change* is used as generally referring to a change in meaning, either of a lexical unit (word or expression) or of a concept (a complex knowledge structure that can encompass one or more lexical units as well as relations among them and with other concepts). Semantic change and other related terms, such as *semantic shift*, *semantic drift*, *concept drift*, *concept shift*, *concept split*, are also introduced and explained in the context used by the authors considered for discussion.

The current study is developed as part of the use case in the humanities (UC4.2.1) carried out within the COST Action *European network for Web-centred linguistic data science (Nexus Linguarum)*, CA18209. [3] The goal of the use case is to create a workflow for the detection of semantic change in multilingual diachronic corpora from the humanities domain, and the representation of the evolution of parallel concepts, derived from these corpora, as LLOD. The intended outcome of UC4.2.1 is a set of diachronic ontologies in several languages and methodological guidelines for generating and publishing this type of knowledge using NLP and Semantic Web technologies. Thus, the paper is organised in eight sections describing the survey methodology and the state-of-the art in data, tools, and methods for NLP and LL(O)D resources that we deem important to a workflow designed for the diachronic analysis and ontological representation of concept evolution. Our main focus is the concept change for humanities research, which involves investigations and data that include a time dimension, but the concepts may also apply to other domains. The various sections will focus on the essential aspects needed to understand the current trends and to build applications for detecting and representing semantic change.

The remainder of this paper is organised as follows. Section 2 presents the methodology applied to build the survey. Section 3 discusses existing theoretical frameworks for tracing different types of semantic

*Corresponding author. E-mail: florentina.armaselu@uni.lu.

[1] We have added parentheses around the word 'open' because although the focus is often on linked data, and in our case linguistic linked data, that has been published with an open license, this is not always the case and linked data may have other types of license.

[2] One exception is [12].

[3] https://nexuslinguarum.eu/.

change. Section 4 presents current LL(O)D formalisms (e.g. RDF, OntoLex-Lemon, OWL-Time) and models for representing diachronic relations. Section 5 is dedicated to existing methods and NLP tools for the exploration and detection of semantic change in large sets of data, e.g. diachronic word embeddings, named entity recognition (NER) and topic modelling. Section 6 presents an overview of methods and NLP tools for (semi-) automatic generation of (diachronic) ontological structures from text corpora. Section 7 provides an overview of the main diachronic LL(O)D repositories from the humanities domain, with particular attention to collections in various languages, and emerging trends in publishing ontologies representing semantic change as LL(O)D data. The paper is concluded by Section 8 where we discuss our findings and future directions.

2. Survey methodology

The motivation of combining DH approaches with semantic technologies is mainly related to the target audiences of the survey. That is, researchers, students, teachers interested in detecting how concepts in a certain domain evolve and how this evolution can be represented via semantic Web and linked data technologies that support the production and dissemination of FAIR data on the Web. Therefore, the paper addresses the study of semantic change and creation of diachronic ontologies in connection with areas in the humanities such as the history of concepts and history of ideas, on the one side, and linguistics, on the other. This topic may be of potential interest to other researchers interested in semantic change detection within a particular domain and its modelling as linked data. Scholars in Semantic Web technologies may be interested as well in such areas of application and further development of the linked data paradigm and the possibilities of integrating diachronic representation of data from the humanities into the LL(O)D cloud in the future.

The scope of the paper covers diachronic corpora that may span more distant or more recent periods in time. Therefore, the article focuses on studies dealing with diachronic variation, that is change over time, but not with synchronic variation, which can refer, for instance, to variation across genre (or register), class, gender or other social category [13], within a given, more limited period of time. The survey also targets the construction of diachronic ontologies that, unlike synchronic ontologies ignoring the historical perspective, allow us to capture the temporal dimension of concepts and investigate gradual semantic changes and concept evolution through time [14].

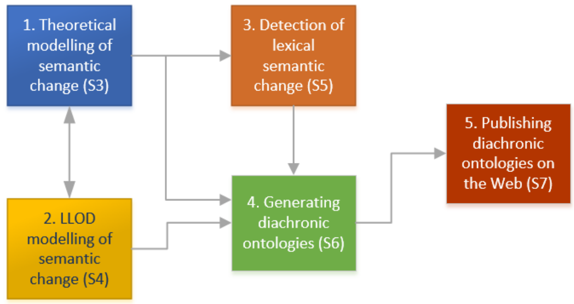

Fig. 1. Generic workflow and related sections

As mentioned above, the survey follows a workflow for detecting and representing semantic change as LL(O)D ontologies, based on diachronic corpora. Figure 1 illustrates the main blocks of such a generic workflow and the possible interconnections among the various areas of research considered relevant for the study. Each block can be mapped onto one of the subsequent sections (referred to as S3 - S7, in Fig. 1). It should be noted that not all of the relationships displayed in the figure are explicitly expressed in the surveyed literature. Some of them represent work in progress or projections of possible developments implied by the intended workflow. For instance, we consider that theoretical modelling of semantic change in diachronic corpora can play an important role in designing the following steps in the workflow, such as LL(O)D modelling, detection of lexical semantic change and ontology generation, and thus, a survey of this area is worth investigating together with the other blocks. Moreover, approaches from the domain of lexical semantic change detection may inform and potentially bring about new perspectives on learning/generating (diachronic) ontologies from unstructured texts, which in turn, connects with existing or prospected means of publishing such ontologies in the LL(O)D cloud.

The methodology consisted, therefore, of three phases: (1) selecting or searching for (recent) surveys or reference works in areas related to the five blocks depicted in Fig. 1; (2) expanding the set by considering relevant references cited in the works collected during the previous phase; (3) refining and balancing the structure of the covered areas and corresponding sections and sub-sections. Phase 1 started with works already known to the authors, as related to their field of research, or resulting from search by keywords such as "semantic change/shift/drift", "history of concepts/ideas", "historical linguistics/semantics", "diachronic/synchronic variation/ontology", "ontology generation/acquisition/extraction/learning". Table 1 summarises the structure and size of the referenced material presented in the survey.

Table 1. Structure and size of the surveyed material

Section	***Related research areas***	***Cited works***
S1, S2	Contextualisation of the topic, survey methodology	14
S3	History of ideas, history of concepts, philosophy, knowledge organisation	9
	Lexical semantics, cognitive linguistics, diachronic lexicology, terminology, pragmatics, discourse analysis	20
S4	The OntoLex-Lemon model	3
	Etymologies as LL(O)D	9
	SW-based modelling of diachronic relations	7
	SW resources for temporal information	4
S5	Overview	20
	NLP Challenges	32
	NER and NEL	24
	Word embeddings	14
	Transformer-based language models	5
	Topic modelling	14
S6	Ontology learning	10
	Diachronic constructs	11
	Generating linked data	7
S7	Diachronic datasets in the LL(O)D cloud, publishing diachronic ontologies as LL(O)D	9
	Total (of which 20 repeated citations)	**192**

3. Theoretical frameworks

Different disciplines (within or applied in the humanities) make use of different interpretations, theoretical notions and approaches in the study of semantic change. In this section, we survey various theoretical frameworks that depart either from knowledge or from language and that can serve the theoretical modelling purposes of block 1 in the generic workflow (Fig. 1). The distinctions between the two lines of enquiry mainly refer to two potentially distinct traditions, one coming from philosophy, history of concepts and history of ideas, the other from linguistics. Although there are no strict demarcations between the two threads and some overlap exists, the first seems to lead to approaches turning to Semantic Web technologies (and the corresponding representation of knowledge, including ontologies), the second to various approaches in corpus-based analysis.

3.1. Knowledge-oriented approaches

Scholars in domains such as history of ideas, history of concepts and philosophy focus on concepts as units of analysis. In his comparative reading of German and English conceptual history, Richter [15] accounts for the distinction between words and concepts in charting the history of political and social concepts, where a concept is understood as a “forming part of a larger structure of meaning, a semantic field, a network of concepts, or as an ideology, or a discourse” (p. 10). Basing his study on three major reference works by 20th-century German-speaking theorists, Richter notes that outlining the history of a concept may sometimes require tracking several words to identify continuities, alterations or innovations, as well as a combination of methodological tools from history, diachronic, and synchronic analysis of language, semasiology, onomasiology, and semantic field theory. He also highlights the importance of sources (e.g. dictionaries, encyclopaedias, political, social, and legal materials, professional handbooks, pamphlets and visual, nonverbal forms of expression, journals, catechisms and almanacs) and procedures to deal with these sources, employed in tracing the history of concepts in a certain domain, as demonstrated by the considered reference works.

Within the framework of intellectual history, Kuukkanen [16] proposes a vocabulary allowing for a more formal description of conceptual change, in response to critiques of Lovejoy’s long-debated notion of “unit-ideas” or “unchangeable concepts”. Assuming that a concept X is composed by two parts, the “core” and the “margin”, underlain by context-unspecific and context-specific features, Kuukkanen describes the core as “something that all instantiations must satisfy in order to be ‘the same concept”’, and the margin as “all the rest of the beliefs that an instantiation of X might have” (p. 367). This paradigm enables us to record a full spectrum of possibilities, from conceptual continuity, implying core stability and different degrees of margin variability, to conceptual replacement, when the core itself is affected by change.

Another type of generic formalisation, combining philosophical standpoints on semantic change, theory of knowledge organisation and Semantic Web technologies, is proposed by Wang et al. [11] who consider that the meaning of a concept can be defined in terms of "intension, extension and labelling applicable in the context of dynamics of semantics" (p. 1). Thus, since reflecting a world in continuous transformation, concepts may also change their meanings. This process, called "concept drift"[4], occurs over time but other kinds of factors, such as location or culture, may be taken into account. The proposal is framed by two "philosophical views" on the change of meaning of a concept over time assuming that: (1) different variants of the same concept can have different meanings (*concept identity* hypothesis); (2) concepts gradually evolve into other concepts that can have almost the same meaning at the next moment in time (*concept morphing* hypothesis). In line with a tradition in philosophy, logic and semiotics going back to Frege's "sense" and "reference" [18] and de Saussure's "signifier" [19], Wang et al. formally describe the meaning of a concept C as a combination of three "aspects": a "set of properties (the intension of C)", a "subset of the universe (the extension of C)", and a "String" (the label) [11, p. 6]. Based on these statements, they develop a system of formal definitions that allows us to detect different forms of conceptual drift, including "concept shift" (where "part of the meaning of a concept shifts to some other concept") and "concept split" (when the "meaning of a concept splits into several new concepts") (pp. 2, 10). Various similarity and distance measures (e.g. Jaccard and Levenshtein) are computed for the three aspects to identify such changes, according to the two philosophical perspectives mentioned above. Within four case studies, the authors apply this framework to different vocabularies and ontologies in SKOS, RDFS, OWL and OBO [5] from the political, encyclopaedic, legal and biomedical domains.

Drawing upon methodologies in history of philosophy, computer science and cognitive psychology, and elaborating on Kuukkanen's and Wang et al.'s formalisations, Betti and Van den Berg [20] devise a model-based approach to the "history of ideas or concept drift (conceptual change and replacement)" (p. 818). The proposed method deems ideas or concepts (used interchangeably in the paper) as models or parts of models, i.e. complex conceptual frameworks. Moreover, it is considered that "concepts are (expressible in language by) (categorematic) terms, and that they are compositional; that is, if complex, they are composed of subconcepts" (p. 813). Arguing that both the *intension* and the *extension* of a concept should be included in the study of concept drift, Betti and Van den Berg identify the former with the core and margin, or meaning, and the latter with the reference. To illustrate their proposal, the authors use a model to represent the concept of "proper science" as a relational structure of fixed conditions (core) containing sub-concepts that can be instantiated differently within a certain category, i.e. of expressions referring to something that can be true, such as 'propositions', 'judgements' or 'thoughts' (margin) (pp. 822 - 824). According to [20], such a model would support the study of the development of ideas by enabling the representation of "concept drift as change in a network of (shifting) relations among subideas" and "fine-grained analyses of conceptual (dis)continuities" (pp. 832 - 833).

Starting with an overview of concept change approaches in different disciplines, such as computer science, sociology, historical linguistics, philosophy, Semantic Web and cognitive science, Fokkens et al. [12] propose an adaption of [16]'s and [11]'s interpretations for modelling semantic change. Unlike [11], [12] argue that only changes in the concept's intension (definitions and associations), provided that the core remains intact, are likely to be understood as concept drift across domains; what belongs to the core being decided by domain experts (oracles). Changes of the core would determine conceptual replacement (following [16]), while changes in the concept's extension (reference) or label (words used to refer to it) are considered related phenomena of semantic change that may or may not be relevant and indicative of concept drift. [12] apply these definitions in an example using context-dependent properties and an RDF representation in Lemon[6]. The authors also draw attention to the fact that making the context of applicability of certain definitions explicit can help in detecting conceptual changes in an ontology and distinguish be-

[4]The term "semantic drift" is also used, although the difference is not explicitly defined. See also the discussion on [17].

[5]SKOS (Simple Knowledge Organization System); RDFS (RDF Schema), RDF (Resource Description Format); OWL (the W3C Web Ontology Language); OBO (Open Biomedical Ontologies).

[6]Lemon (the Lexicon Model for Ontologies).

tween changes in the world, that need to be formally tracked, and changes due to corrections of inadequate or inaccurate representations. However, obtaining the required information for the former case appears to be a challenging task, a possible path of investigation mentioned in the paper referring to recent advances in distributional semantics that can be effective in capturing semantic change from texts.

A different interpretation is offered by Stavropoulos et al. [17] through a background study intended to describe the usage of terms such as *semantic change*, *semantic drift* and *concept drift* in relation to ontology change over time and according to different approaches in the field. Thus, from the perspective of evolving semantics and Semantic Web, the authors frame semantic change as a "phenomenon of change in the meaning of concepts within knowledge representation models". More precisely, semantic change denotes "extensive revisions of a single ontology or the differences between two ontologies and can, therefore, be associated with versioning" (p. 1). Within the same framework, they define semantic drift as referring to the gradual change either of the features of ontology concepts, when their knowledge domain evolves, or of their semantic value, as it is perceived by a relevant user community. Further distinction are drawn between *intrinsic* and *extrinsic* semantic drift, depending on the type of change in the concept's semantic value. That is, in respect to other concepts within the ontology or to the corresponding real world object referred by it. Originated from the field of incremental concept learning [21] and adapted to the new challenges of the Semantic Web dynamics [22], concept drift is described in [17, p. 3] as a "change in the meaning of a concept over time, possibly also across locations or cultures, etc.". Following [11], three types of concept drifts are identified as operating at the level of *label*, *intension* and *extension*. Stavropoulos et al. transfer this type of formalisation to measure semantic drift in a dataset from the *Software-based Art* domain ontology, via different similarity measures for sets and strings, by comparing each selected concept with all the concepts of the next version of the ontology and iterating across a decade. The two terms, semantic drift and concept drift, initially emerged from different fields but according to [17] an increasing number of studies show a tendency to apply notions and techniques from a field to the other.

3.2. Language-oriented approaches

Scholars from computational semantics employ a slightly different terminology than scholars from history of ideas, history of concepts and philosophy. Kutuzov et al. [8], for example, describe the evolution of word meaning over time in terms of "lexical semantic shifts" or "semantic change", and identify two classes of semantic shifts: "linguistic drifts (slow and regular changes in core meaning of words) and cultural shifts (culturally determined changes in associations of a given word)" (p. 1385).

Disciplines from more traditional linguistics-related areas provide other types of theoretical bases and terminologies to research in semantic change and concept evolution. For instance, Kvastad [23] underlines the distinction made in semantics between concept and ideas, on one side, and terms, words and expressions, on the other side, where a "concept or idea is the meaning which a term, word, statement, or act expresses" (p. 158). Kvastad also proposes a set of methods bridging the field of semantics and the study of the history of ideas. Such approaches include synonymity, subsumption and occurrence analysis allowing the historians of ideas to trace and interpret concepts on a systematic basis within different contexts, authors, works and periods of time. Other semantic devices listed by the author can be used to define and detect ambiguity in communication between the author and the reader, formalise precision in interpretation or track agreement and disagreement in the process of communication and discussion ranging over centuries.

Along a historical timeline, spanning from the middle of the 19th-century to 2009, Geeraerts [24] presents the major traditions in the linguistics field of lexical semantics, with a view on the theoretical and methodological relationships among five theoretical frameworks: historical-philological semantics, structuralist semantics, generativist semantics, neostructuralist semantics and cognitive semantics. While focusing on the description of these theoretical frameworks and their interconnections in terms of affinity, elaboration and mutual opposition, the book also provides an overview on the mechanisms of semantic change within these different areas of study. The main classifications of semantic change resulted from historical-philological semantics include on one hand, the semasiological mechanisms (*meaning*-related) that "involve the creation of new readings within the range of application of an existing lexical item", with semasiological innovations endowing existing words with new meanings. On the other hand, the onomasiological (or "lexicogenetic") mechanisms (*naming*-related) "involve changes through which a concept, regardless of whether or not it has previously been lexicalised, comes to be expressed by a new or alternative lexical item", with onomasiological innovations coupling "concepts to words in a way that is not yet part of the lexical inventory of the language" (p. 26). Further distinctions within the first category refer to lexical-semantic changes such as specialisation and generalisation, or metonymy and metaphor. On the other hand, the second category is related to the process of word formation that implies devices such as morphological rules for derivation and composition, transformation through clipping or blending, borrowing from other languages or onomatopoeia-based development. Geeraerts also points out the general orientation of historical-philological semantics as diachronic and predominantly semasiological rather than onomasiological, with a focus on the change of meaning understood as a result of psychological processes, and an "emphasis on shifts of conventional meaning" and thus an empirical basis consisting "primarily of lexical uses as may be found in dictionaries" (p. 43). In this sense, historical-philological semantics links up with lexicography, etymology and history of ideas ("meanings are ideas") (p. 9). Moreover, the author distinguishes three main perspectives: *structural* that looks at the "interrelation of [linguistic] signs" (sign-sign relationship), *pragmatic* that considers the "relation between the sign and the context of use, including the language user" (sign–use(r) relationship), and *referential* that delineates the "relation between the sign and the world" (sign–object relationship). According to [24], the evolution of lexical semantics (and implicitly of the way meaning and semantic change are reflected upon) can be characterised therefore by an oscillation along these three dimensions. A historical-philological stage dominated by the referential and pragmatic perspective, a structuralist phase centred on structural, sign–sign relations, an intermediate position shaped by generativist and neostructuralist approaches, and a current cognitive stance that recontextualises semantics within the referential and pragmatic standpoint and displays a certain affinity with usage-based approaches such as distributional analysis of corpus data (pp. 278 - 279, 285).

In cognitive linguistics and diachronic lexicology Grondelaers et al. [25] also identify that semantic change could be approached by applying two different perspectives – onomasiological and semasiological. The onomasiological approach focuses on the referent and studies diachronically the representations of the referent, whereas the semasiological approach investigates the linguistic expression by researching diachronically the variation of the objects identified by the linguistic expression under the investigation. There is a tendency to apply the semasiological approach in computational semantic change research because it relies on words or phrases extracted from the datasets; however, the extraction of concept representations from linguistic data poses certain challenges and requires either semi-automatically or automatically learning ontologies to trace concept drift or change as it was discussed above.

In other fields, such as terminology, semasiological and onomasiological approaches may encompass either a concept- or a term-oriented perspective [26, 27]. Other standpoints, framed for instance in a sociocognitive context, attempt to take into account both the principles of stability, univocity of "one form for one meaning" and synchronic term-concept relationship from traditional terminology, and the need for understanding and interpreting the world and language in their dynamics as they change over time, and for applying more flexible tools when analysing semantic change in a specialised domain, such as prototype theory [28, pp. 126, 130)].

Diachronic change in the layer of pragmatics is a specific task requiring special endeavor as it is context specific. While analysing diachronic change of discourse markers, first it should be stressed that the notion of discourse marker was introduced by Schiffrin [29] and the author considered phrases such as 'I think' a discourse marker performing the function of discourse management deictically "either point[ing] backward in the text, forward, or in both directions". Fraser [30] provided a taxonomy of pragmatic markers drawn from syntactic classes of conjunctions, adverbials and prepositional phrases followed by Aijmer [31] suggesting that 'I think' is a "modal particle". Over the last few decades the research on discourse markers has developed into a considerable and independent field accepting the term of discourse markers [32–34]

Another point that deals with the manual analysis of diachronic change of discourse markers, e.g., Waltereit and Detges [35] analysed the development of the Spanish discourse marker *bien* derived from the Latin manner adverb *bene* ('well') and showed that the functional difference between discourse markers and modal particles can be related to different diachronic pathways. Currently, corpus-driven automatic analysis is acquiring the impetus, e.g. Stvan and Smith [36] use corpus analysis relating early 20th-century American texts with modern TV shows to research diachronic change in the discourse markers ‘why’ and ‘say’ in American English. However, there are still challenges analysing diachronic change on the pragmatic layer as there is a need for a move from queries based on individual words towards larger linguistic units and pieces of text.

In addition to linguistic approach focusing on text linguistics and pragmatics, discourse analysis in a broad sense studies naturally occurring language referring to socio-related textual characteristics in humanities and social sciences. According to Foucault, one of the key theorists of the discourse analysis, the term „discourse“ refers to institutionalized patterns and disciplinary structures concerned with the connection of knowledge and power [37]. Discourse analysis approaches language as a means of social interaction and is related to the social contexts embedding the discourse. Within this framework, the discourse-historical approach (DHA) is of particular interest, as part of the broader field of critical discourse analysis (CDA) that investigates "language use beyond the sentence level" and other "forms of meaning-making such as visuals and sounds" as elements in the "(re)production of society via semiosis" [38]. Thus, based on the principle of "triangulation", DHA takes into account a variety of data, methods, theories and background information to analyse the historical dimension of discursive events and the ways in which specific discourse genres are subject to diachronic change. Recent studies on linguistic change using diachronic corpora and a combination of computational methods, such as word embedding, and discourse-based approaches argue that a discourse-historical angle can provide a better understanding of the interrelations between language and social, cultural and historical factors, and their change over time [39, 40].

4. LOD formalisms

Having given an overview of different theoretical perspectives on semantic change across numerous disciplines in the (digital) humanities-related areas, in the current section we will look at how some of these perspectives can be modelled as linked data. In particular, we enquire about possible modalities for formally representing the evolution of word meanings and their related concepts over time within a LL(O)D and Semantic Web framework (also in connection to block 2, Fig. 1). In Section 4.1, we will look at the OntoLex-Lemon model for representing lexicon-ontologies as linked data. This model is useful for representing the relationship between a lexicon and a set of concepts, something that is relevant for both knowledge-oriented and language-oriented approaches mentioned in Section 3. Next, in Section 4.2, we look at the representation of etymologies or word-histories in linked data as these are particularly useful in language-oriented approaches to semantic change. Afterwards, in Section 4.4 we look at how to explicitly represent diachronic relations in RDF; this is useful for any situation in which we have to model dynamic information and is relevant to both of the general approaches in Section 3 and is not limited only to linked data. Finally, we look at resources for representing temporal information in linked data, in Section 4.4.

4.1. The OntoLex-Lemon model

OntoLex-Lemon [41] is the most widely used model for publishing lexicons as linked data. For what regards its modelling of the semantics of words, it represents the meaning of any given lexical entry “by pointing to the ontological concept that captures or represents its meaning” [7]. In OntoLex-Lemon, the class LexicalSense is defined as “[representing] the lexical meaning of a lexical entry when interpreted as referring to the corresponding ontology element”, that is “a reification of a pair of a uniquely determined lexical entry and a uniquely determined ontology entity it refers to”. Moreover, the object property sense is defined in the W3C Community Report as “[relating] a lexical entry to one of its lexical senses” and the object propertyreference as “[relating] a lexical sense to an ontological predicate that represents the denotation of the corresponding lexical entry”. See Figure 2 for a schematic representation of the OntoLex-Lemon core. Another property that is relevant to the modelling of lexical meaning is denotes which is equivalent to the property chain sense o reference [8]. In addition, the Usage class allows us to describe sense usages of individuals of LexicalSense.

[7] Lexicon Model for Ontologies: Community Report, 10 May 2016 (w3.org) https://www.w3.org/2016/05/ontolex/#semantics

[8] Here o stands for the relation composition operator, i.e., $(a, b) \in RoS \Leftrightarrow \exists c.(a, c) \in R\&(c, b) \in S$

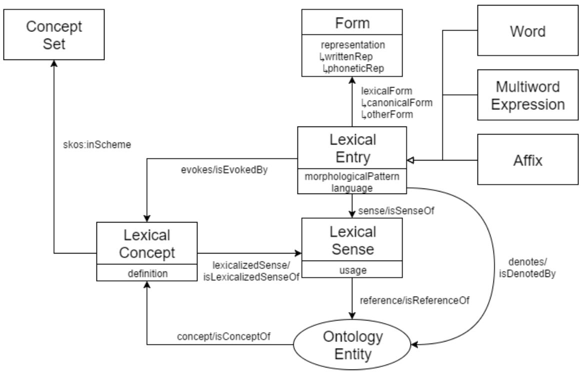

Fig. 2. OntoLex-Lemon core model

OntoLex-Lemon also allows users the possibility of modelling *usage* conditions on a lexical sense – conditions that reflect pragmatic constraints on word meaning such as those which concern register – via the (appropriately named) object property usage [9]. The use of this property is intended to complement the lexical sense rather than to replace it.

To summarise, then, OntoLex-Lemon offers users a model for representing the relationship between a lexical sense and an ontological entity in linked data. The relationship between lexical and conceptual aspects, or more broadly speaking linguistic and conceptual aspects of meaning [10], is important for many of the approaches listed in Section 3: both the knowledge-oriented approaches described in Section 3.1 such as those of Richter as well as the language-oriented approaches of Section 3.2. Note that the work of [12] described above in Section 3.1 is already based on *lemon*, the immediate pre-cursor of OntoLex-Lemon.

Another OntoLex-Lemon class for modelling meaning is LexicalConcept. This is defined as "a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses" in the OntoLex-Lemon guidelines [11] It is related to LexicalEntry via the evokes class which relates a lexical entry to a "mental concept that speakers of a language might associate when hearing [the entry]". From this definition then a lexical entry for the word *grape* could be related via evokes to the concept of 'wine' or 'harvest' or specific geographical regions such as Burgundy or Concord. This might be useful in tracing the different associations and related concepts that a word picks up over time, while sense and reference are used to look at the core intensional and extensional meanings of the same words.

Work on a Frequency, Attestation and Corpus Information module (FrAC) for OntoLex-Lemon is underway in the OntoLex W3C group [43]. This module, once finished, will enable the addition of corpus-related information to lexical senses, including information pertaining to word embeddings.

4.2. *Representing etymologies and sense shifts in LL(O)D*

One important source of information on semantic shifts are etymologies. These are defined as word histories and include descriptions of both the linguistic drifts and cultural shifts described by Kutuzov et al. and discussed in Section 3.2 as well as the other (language-related) approaches discussed in that section. They can be used in some of the knowledge-oriented approaches mentioned in Section 3.1 such as that of Richter. As well as being a *source* of semantic change information, etymologies can also be used to encode and to make accessible semantic change information in lexical resources in a standardised way; we can do this by making use of and extending existing linked data vocabularies as we will see in this section.

Current work in modelling etymology in LL(O)D was preceded and influenced by similar work in related standards such as the Text Encoding Initiative (TEI) and the Lexical Markup Framework (LMF). This includes notably Salmon-Alt's LMF-based approach to representing etymologies in lexicons [44], as well as Bowers and Romary's [45] work which builds on already existing TEI provisions for encoding etymologies in order to propose a *deep* encoding of etymological information in TEI. In the latter case, the authors' approach entailed enabling an enhanced structuring of lexical data that would allow for the identification of, for instance, etymons and cognates in a TEI entry, as well as the specification of different varieties of etymological change. This latter work also coincides with the development, currently in progress, of an etymological extension of LMF by the International Standards Organization working group ISO/TC 37/SC 4/WG 4 [46], see also [47] for examples of LMF encoding from a Portuguese dictionary, the *Grande Dicionário Houaiss da Língua Portuguesa*.

[9] https://www.w3.org/2016/05/ontolex/#usage

[10] Note that ontologies are usually described as *conceptualisations* and of consisting of *concepts* [42] which makes them an ideal prerequisite for modelling conceptual shift.

[11] https://www.w3.org/2016/05/ontolex/#lexical-concept-class

Work on the representation of etymologies in RDF instead includes de Melo's [48] work on *Etymological WordNet*, as well as Chiarcos et al's [49] definition of a minimal extension of the *lemon* model with two new properties `cognate` and the transitive `derivedFrom` for representing etymological relationships. Khan [50] defines an extension of OntoLex-Lemon that, like [45] attempts to facilitate a more detailed encoding of etymological information. Notably, this extension reifies the notion of etymology defining individuals of the Etymology class as containers for an ordered series of EtymologicalLink individuals. The latter class is once again a reification, this time of the notion of an etymological link. These etymological link objects connect together Etymon individuals and (OntoLex) Lexical Entries or indeed any other kinds of lexical element that can have an etymology. We can subtype etymological links in order to represent sense shifts within the same lexical entry. Other work specifically on the modelling of sense shift in LL(O)D includes the modelling of semantic shift in Old English emotion terms in [51] in which semantic shifts are reified and linked to elements in a taxonomy of metonymy and metaphor which describe the conceptual structure of these shifts.

Etymological datasets in LL(O)D include the Latin-based etymological lexicon published as part of the LiLa project and described in [52].

4.3. Representing diachronic relations

We have so far looked at ways of representing information about lexicons and the concepts which they lexicalise in RDF, and which are salient for both knowledge-oriented and language-oriented approaches. However, as argued by [53], in order to be able to represent changes in the meaning of concepts, as well as the concepts themselves within the framework of the OntoLex-Lemon model, it would be useful to be able to add temporal parameters to (at least) the properties sense or reference, as well as possibly the *evokes* property. We refer to such properties or relations that can change with time as *fluents*. Due to a well known expressive limitation of the RDF framework, it is not possible to add a temporal parameter to a binary properties. In order to remedy this state of affairs we can either extend RDF or use a number of suggested ontology design patterns in order to stay within the expressive constraints of RDF.

An example of the first strategy is described in [54] where Rizzolo et al. present a formal "RDF-like model" for concept evolution. This is based both on the idea of temporal knowledge bases, in which temporal intervals or lifespans are associated with resources as well as new relations for expressing parthood and causality between concepts. These relations underpin the authors' representation of concept evolution via specialised terms. Finally, they present a special extension of SPARQL based on their new framework and which permits the querying of temporal databases for questions relating to the evolution of a concept over a time period. In [55], Gutierrez et al. propose an extension of RDF which permits temporal reasoning and which describes so-called temporal RDF graphs. They present a syntax, semantics as well as an inference system for this new extension[12], as well as a new temporal query language. Another more recent solution which is still under active development at the time of the writing of this paper is RDF* (also known as RDF-star) [13]. In RDF*, triples can be embedded in and therefore described by other triples. This means for instance that a relationship such as sense can be associated with temporal properties which delimit its temporal validity. In terms of the second solution, there are numerous design patterns for adding temporal information to RDF and permitting temporal reasoning over RDF graphs without adding extra constructs to the language. We will look very briefly at a few of the most prominent of these, however see [56] for a more detailed survey.

The first pattern we will look at, proposed by the W3C as a general strategy for representing relations with an arity greater than 2, is to reify the relation in question, that is turn it into an object. According to this pattern we could turn OntoLex-Lemon sense and reference relations into objects. This pattern has the disadvantage of being too prolix and creating a profusion of new objects, it also means that we cannot use certain OWL constructs for reasoning (see [57] for more details).

Other prominent patterns take the *perdurantist* approach by modelling entities as having temporal parts, as well as (for physical objects) physical parts. Perhaps the most influential of these is the Welty-Fikes pattern introduced in [57] where fluents are represented as holding between temporal parts of entities rather than the entities themselves. For instance, the OntoLex-Lemon property sense would hold between

[12] They are able to show that their entailment for temporal RDF graphs does not lead to an asymptotic increase in complexity.

[13] A draft of the specification can be found at this link: https://w3c.github.io/rdf-star/cg-spec/editors_draft.html

temporal parts of LexicalSense individuals rather than the individuals themselves. The Welty-Fikes pattern is much less verbose than the first pattern, and also allows us to use the OWL constructs alluded to in the last paragraph. However the fact that the Welty-Fikes pattern constrains us into redefining fluent properties as holding between temporal parts rather than between the original entities (so sense, or the temporal version, would no longer have the OntoLex-Lemon classes LexicalEntry as a domain and LexicalSense as a range) could be seen as a serious disadvantage. A simplification to the Welty-Fikes pattern is proposed in [58] in which "what has been an entity becomes a time slice". This implies that fluents hold between perdurants, that is entities with a temporal extent, but these can be, in our example, lexical entries and senses. This is the approach which was taken in [59] in order to model dynamic lexical information, and where lexical entries and senses (among other OntoLex-Lemon elements) were given temporal extents.

4.4. OWL-Time ontology and other Semantic Web resources for temporal information

The most well known linked data resource for encoding temporal information is the OWL-Time ontology [60]; as of March 2020 it is a W3C Candidate Recommendation. OWL-Time allows for the encoding of temporal facts in RDF, both according to the Gregorian calendar as well as other temporal reference systems, including alternative historical and religious calendars. It includes classes representing time instants and time intervals as well as provision for representing topological relationships among intervals and instants and in particular those included in the Allen temporal interval algebra [61]. This allows for reasoning to be carried out over temporal data that uses the Allen properties, in conjunction with an appropriate set of OWL axioms and SWRL rules, such as those described in [62].

Other useful resources that should be mentioned here are PeriodO[14], an RDF-based gazetteer of temporal periods which are salient for work in archaeology, history and art-history [63] and LODE, *an ontology for Linking Open Descriptions of Events*[15]. These resources are useful both for approaches which deal specifically with linguistic linked data as well as those which deal with shifts in concepts over time more generally.

[14]https://perio.do/en/

[15]https://linkedevents.org/ontology/

5. NLP for detecting lexical semantic change

Given the possibilities described above for modelling semantic change via LL(O)D formalisms, we will address the question of automatically capturing such changes in word meaning (block 3, Fig. 1) by analysing diachronic corpora available in electronic format. This section draws an overview of existing methods and NLP tools for the exploration and detection of lexical semantic change in large sets of data, e.g. related to diachronic word embeddings, named entity recognition (NER) and topic modelling.

5.1. Overview

The past decade has seen a growing interest in computational methods for lexical semantic change detection. This has spanned across different communities, including NLP and computational linguistics, information retrieval, digital humanities and computational social sciences. A number of different approaches have been proposed, ranging from topic-based models [64–66], to graph-based models [67, 68], and word embeddings [10, 69–75]. [7], [6], and [8] provide comprehensive surveys of this research until 2018. Since then, this field has advanced even further [76–79].

In spite of this rapid growth, it was only in 2020 that the first standard evaluation task and data were created. [9] present the results of the first SemEval shared task on *unsupervised lexical semantic change detection*, which represents the current NLP state of the art in this field. Thirty-three teams participated in the shared task, submitting 186 systems in total. These systems consist in a representation of the semantics of words from the input diachronic corpus, which is normally split into subcorpora covering different time intervals. The majority of the methods proposed rely on embedding technologies, including type embeddings (i.e. average embeddings representing a word type) and token embeddings (i.e. contextualised embeddings for each token). Once the semantic representations have been built, a method for aligning these representations over the temporal sub-corpora is needed. The alignment techniques used include orthogonal Procrustes [10], vector initialisation [69] and temporal referencing [78]. Finally, in order to detect any significant shift which can be interpreted as semantic change, the change between the representations of the same word over time needs to be measured. The change measures typically used include distances based on cosine and local neighbours, Kullback-Leibler divergence, mean/standard deviation of co-occurrence vectors, or cluster frequency. The systems which participated in the shared task were evaluated on manually-annotated gold standards for four languages (English, German, Latin and Swedish) and two sub-tasks, both aimed at detecting lexical semantic change between two time periods: given a list of words, the binary classification sub-task aimed at detecting which words lost or gained senses between the two time periods, while the ranking sub-task consisted in ranking the words according to their degree of semantic change between the two time periods. The best-performing systems all use type embedding models, although the quality of the results differs depending on the language. Averaging over all four languages, the best result had an accuracy of 0.687 for sub-task 1 and a Spearman correlation coefficient of 0.527 for sub-task 2.

5.2. NLP Challenges

Detecting lexical semantic change via NLP may imply a series of challenges, related to the digitisation, preparation and processing of data, as discussed below.

Applying NLP tools, such as POS taggers, syntactic parsers, and named entity recognisers to historical texts is difficult, because most existing NLP tools are developed for modern languages [80, 81]. A historical language often differs significantly from its modern counterpart. The two often have different linguistic aspects, such as lexicon, morphology, syntax, and semantics which make a naive use of these tools problematic [82, 83]. One of the most prevalent differences is spelling variation. The detection of spelling variants is an essential preliminary step for identifying lexical semantic change. A frequently suggested solution for the spelling variation issue is normalisation. Normalisation is generally described as the mapping of historical variant spellings into a single, contemporary "normal form".

Recently, Bollmann [84] systematically reviewed automatic historical text normalisation. Bollmann divided the research data into six conceptual or methodical approaches. In the first approach, each historical variant is checked in a compiled list that maps its expected normalisation. Although this method does not generalise patterns for variants not included in the list, it has proved highly successful as a component of several other normalisation systems [85, 86]. The second approach is rule-based. The rule-based approach aims to encode regularities in the form of substitution rules in spelling variations, usually including context information to distinguish between different character uses. This approach has been adopted to various languages including German [87], Basque, Spanish [88], Slovene [89], and Polish [90]. The third approach is based on editing distance measures. Distance measures are used to compare historical variants to modern lexicon entries [86, 91, 92]. Normalisation systems often combine several of these three approaches [85, 92–94]. The fourth approach is statistical. The statistical approach models normalisation as a probability optimisation task, maximising the probability that a certain modern word is the normalisation of a given historical word. The statistical approach has been applied as a noisy channel model [89, 95], but more commonly as character-based statistical machine translation (CSMT) [96–98], where the historical word is "translated" as a sequence of characters. The fifth approach is based on neural network architectures, where the encoder–decoder model with recurrent layers is the most common [99–103]. The encoder–decoder model is the logical neural counterpart of the CSMT model. Other works modelled the normalisation task as a sequence labelling problem and applied long short-term memory networks (LSTM) neural networks [104, 105]. Convolutional networks were also used for lemmatisation [106]. In the sixth approach Bollmann [84] included models that use context from the surrounding tokens to perform normalisation [107, 108]. Bollmann [84] also compares and analyses the performance of three freely available tools that cover all types of proposed normalisation approaches on eight languages. The datasets and scripts are publicly available.

Other studies in detecting lexical semantic change pointed out different types of challenges. For instance, in their analysis of markers of semantic change and leadership in semantic innovation using diachronic word embeddings and two corpora containing scientific articles and legal opinions from 20 and 18 century to present, [109] reported difficulties posed by names and abbreviations in identifying genuine candidates of semantic innovations. They applied a series of post-processing heuristics to alleviate these problems, by training a feed-forward neural network and using a pre-trained tagger to label names and proper nouns or to detect abbreviations under a certain frequency threshold, and discarding them from the list of candidates.

[110] addressed the scalability and interpretability issues observed in semantic change detection with clustering of all word's contextual embeddings for

large datasets, mainly related to high memory consumption and computation time. The authors used a pre-trained BERT model (see Section 5.5) to detect word usage change in a set of multilingual corpora (in German, English, Latin and Swedish) of COVID-19 news from January to April 2020. To improve scalability, they limited the number of contextual embeddings kept in memory for a given word and time slice by merging highly similar vectors. The most changing words were identified according to divergence and distance measures of usage computed between successive time slices. The most discriminating items from the clusters of usage corresponding to these words were then used by the researchers and domain experts in the interpretation of results.

Another type of challenge is that of assessing the impact of OCR (Optical Character Recognition) quality on downstream NLP tasks, including the combined effects of time, linguistic change and OCR quality when using tools trained on contemporary languages to analyse historical corpora. [111] performed a large-scale analysis of the impact of OCR errors on NLP applications, such as sentence segmentation, named-entity recognition (NER), dependency parsing and topic modelling. They used datasets drawn from historical newspapers collections and based their tests and evaluation on OCR'd and human-corrected versions of the same texts. Their results showed that the performance of the examined NLP tasks was affected to various degrees, with NER progressively degrading and topic modelling diverging from the "ground truth", with the decrease of OCR quality. The study demonstrated that the effects of OCR errors on this type of applications are still not enough understood, and highlighted the importance of rigorous heuristics for measuring OCR quality, especially when heritage documents and a temporal dimension are involved.

5.3. *Named-entity recognition and named-entity linking*

Named-entity recognition (NER) and named-entity linking (NEL) which allow organisations to enrich their collections with semantic information have increasingly been embraced by the digital humanities (DH) community. For many NLP-based systems, identifying named-entity changes is crucial since failure to know various names referring to the same entity greatly affects their efficiency. Temporal NER has been mostly studied in the context of historical corpora. Various NER approaches have been applied to historical texts including early rule-based approaches [112–114] through unsupervised statistical approaches [115], conventional machine learning approaches [116–118] and to deep learning approaches [119–123]. Named-entity disambiguation (NED) was also investigated and Agarwal et al. [124] introduced the first time-aware method for NED of diachronic corpora.

Different eras, domains, and typologies have been investigated, so comparing different systems or algorithms is difficult. Thus, [125] recently introduced the first edition of HIPE (Identifying Historical People, Places and other Entities), a pioneering shared task dedicated to the evaluation of named entity processing on historical newspapers in French, German and English [126]. One of its subtasks is Named Entity Linking (NEL). This subtask includes the linkage of the named entity to a particular referent in the knowledge base (KB) (Wikidata) or a NEL node if the entity is not included in the base.

Traditionally, NEL has been addressed in two main approaches: text similarity-based and graph-based. Both of these approaches were adapted to historical domains mostly as 'of-the-shelf' NEL systems. While some of the previous works perform NEL using the KB unique ids [126, 127], other works use LL(O)D formalisms [128–131]. One of the aims of the HIPE shared task was to encourage the application of neural-based approaches for NER which has not yet been applied to historical texts. This aim was achieved successfully. Teams have experimented with various entity embeddings, including classical type-level word embeddings and contextualised embeddings, such as BERT (see Section 5.5). The manual annotation guidelines of the HIPE corpus were derived from the Quaero annotation guide [132] and thus, the HIPE corpus mostly remains compatible with the NewsEye project's NE Finnish, French, German, and Swedish datasets [16]. Pontes et al. [133] analysed the performance of various NEL methods on these two multilingual historical corpora and suggested multiple strategies for alleviating the effect of historical data problems on NEL. In this respect, they pointed out the variations in orthographic and grammatical rules, and the fact that names of persons, organisations, and places could have significantly changed over time. [133] also mentioned potential avenues for further research and applications in this area. This may include

[16]https://www.newseye.eu/.

the use of entity linking in historical documents to improve the coverage and relevance of historical entities within knowledge bases, the adaptation of the entity linking approaches to automatically generate ontologies for historical data, and the use of diachronic embeddings to deal with named entities whose name have changed through the time.

Social media communication platforms like Twitter, through informal, colloquial and non-standard language, has led to major changes in the character of written languages. Therefore, in recent years, there has been research interest in NER for social media diachronic corpora. Rijhwani and Preoţiuc-Pietro [134] introduced a new dataset of 12,000 English tweets annotated with named entities. They examined and offered strategies for improving the utilisation of temporally-diverse training data, focused on NER. They empirically illustrated how temporal drift affects performance and how time information in documents can be leveraged to achieve better models.

5.4. *Word embeddings*

The common approach for lexical semantic change detection is based on semantic vector spaces meaning representations. Each term is represented as two vectors representing its co-occurring statistics at various eras. The semantic change is usually calculated by distance metric (e.g. cosine), or by differences in contextual dispersion between the two vectors.

Previously, most of the methods for lexical semantic change detection built co-occurrence matrices [135–137]. While in some cases, high-dimensional sparse matrices were used, in other cases, the dimensions of the matrices were reduced mainly using singular value decomposition (SVD) [138]. Yet, in the last decade, with the development of neural networks, the word embedding approach commonly replaced the mathematical approaches for dimensional reduction.

Word embedding is a collective name for neural network based approaches in which words are embedded into a low dimensional space. They are used as a lexical representation for textual data, where words with a similar meaning have similar representation [139–142]. Although these representations have been used successfully for many natural language pre-processing and understanding tasks, they cannot deal with the semantic drift that appears with the change of meaning over time if they are not specifically trained for this task.

In [143], a new unsupervised model for learning condition-specific embeddings is presented, which encapsulates the word's meaning whilst taking into account temporal-spatial information. The model is evaluated using the degree of semantic change, the discovery of semantic change, and the semantic equivalence across conditions. The experimental results show that the model captures the language evolution across both time and location, thus making the embedding model sensitive to temporal-spatial information.

Another word embeddings approach for tracing the dynamics of change of conceptual semantic relationships in a large diachronic scientific corpus is proposed in [144]. The authors focus on the increasing domain-specific terminology emerging from scientific fields. Thus, they propose to use hyperbolic embeddings [145] to map partial graphs into low dimensional, continuous hierarchical spaces, making more explicit the latent structure of the input. Using this approach, the authors manage to build diachronic semantic hyperspaces for four scientific topics (i.e., chemistry, physiology, botany, and astronomy) over a large historical English corpus stretching for 200 years. The experiments show that the resulting spaces present the characters of a growing hierarchisation of concepts, both in terms of inner structure and in terms of light comparison with contemporary semantic resources, i.e., WordNet.

To deal with the evolution of word representations through time, the authors in [146] propose three LSTM-based sequence to sequence (Seq2Seq) models (i.e., a word representation autoencoder, a future word representation decoder, and a hybrid approach combining the autoencoder and decoder) that measure the level of semantic change of a word by tracking its evolution through time in a sequential manner. Words are represented using the word2vec skip-gram model [139]. The level of semantic change of a word is evaluated using the average cosine similarity between the actual and the predicted word representations through time. The experiments show that hybrid approach yields the most stable results. The paper concludes that the performance of the models increases alongside the duration of the time period studied.

Word embeddings are also used to capture synthetic distortions in textual corpora. In [147], the authors propose a new method to determine paradigmatic (i.e., a term can be replaced by a word) and syntagmatic associations (i.e., the co-occurrence of terms) shifts. The study employs three real-world datasets, i.e., Reddit, Amazon, and Wikipedia, with texts, collected between 1996-2018 for the experiments. The analysis concludes that Local Neighborhood [148], which detects shifts via the k nearest neighbors, is sensitive to paradigmatic shifts while the Global Semantic Displacement [148], which detects shifts within word co-occurrence using the cosine similarity of embeddings, is sensitive to syntagmatic shifts in word embeddings. Furtheremore, the experimental results show that words undergo paradigmatic and syntagmatic shifts both separately and simultaneously.

5.5. *Transformer-based language models*

The current state of the art in word representation for multiple well-known NLP tasks is established by transformer-based pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) [149], ELMo [150] and XLNet [151]. Recently, transformers were also used in lexical semantic change tasks. In paper [152], the authors present one of the first unsupervised approaches to lexical-semantic change that utilise a transformer model. Their solution exploits the BERT transformer model to obtain contextualised word representations, compute usage representations for each occurrence of these words, and measure their semantic shifts along time. For evaluation, the authors utilise a large diachronic English corpus that covers two centuries of language use. The authors provide an in-depth analysis of the proposed model, proving that it captures a range of synchronic, e.g., syntactic functions, literal and metaphorical usage, and diachronic linguistic aspects. In paper [153], different clustering methods are used on contextualised BERT word embeddings to quantify the level of semantic shift for target words in four languages, i.e., English, Latin, German, Swedish. The proposed solutions outperform the baselines based on normalised frequency difference or cosine distance methods.

5.6. *Topic modelling*

Topic modelling is another category of methods proposed for the study of semantic change. Topic modelling often refers to latent Dirichlet allocation (LDA) [154], a probabilistic technique for modelling a corpus by representing each document as a mixture of topics and each topic as a distribution over words. LDA is referred to either as an element of comparison or as a basis for further extensions that take into account the temporal dimension of word meaning evolution. Frermann and Lapata [66] draw ideas from such an extension, the dynamic topic modelling approach [155], to build a dynamic Bayesian model of Sense ChANge (SCAN) that defines word meaning as a set of senses tracked over a sequence of contiguous time intervals. In this model, senses are expressed as a probability distribution over words, and given a word, its senses are inferred for each time interval. According to [66], SCAN is able to capture the evolution of a word's meaning over time and detect the emergence of new senses, sense prevalence variation or changes within individual senses such as meaning extension, shift, or modification. Fremann and Lapata validate their findings against WordNet and evaluate the performance of their system on the SemEval-2015 benchmark datasets released as part of the *diachronic text evaluation* exercise.

Pölitz et al. [156] compare the standard LDA [154] with the continuous time topic model [157] (called "topics over time LDA" in the paper), for the task of word sense induction (WSI) intended to automatically find possible meanings of words in large textual datasets. The method uses lists of key words in context (KWIC) as documents, and is applied to two corpora: the dictionary of the German language (DWDS) core corpus of the 20th century and the newspaper corpus Die Zeit covering the issues of the German weekly newspaper from 1946 to 2009. The paper concludes that standard LDA can be used, to a certain degree, to identify novel meanings, while topics over time LDA can make clearer distinctions between senses but sometimes may result in too strict representations of the meaning evolution.

[64, 65] apply the hierarchical Dirichlet process technique [158], a non-parametric variant of LDA, to detect word senses that are not attested in a reference corpus and to identify novel senses found in a corpus but not captured in a word sense inventory. The two studies include experiments with various datasets, such as selections from the BNC corpus (British English from the late 20th-century), ukWaC Web corpus (built from the .uk domain in 2007), SiBol/Port collection (texts from several British newspapers from 1993, 2005, and 2010) and domain-specific corpora such as sports and finance. Another example is [159] that applies topic modelling to the corpus of Hartlib Papers, a multilingual collection of correspondence and other papers of Samuel Hartlib (c.1600-1662) spanning the period from 1620 to 1662, to identify changes in the topics discussed in the letters. They then experimented

with using topic modelling to detect semantic change, following the method developed in [160].

Based on these overviews and state of the art, we can say that automatic lexical semantic change detection is not yet a solved task in NLP, but a good amount of progress has been achieved and a great variety of systems have been developed and tested, paving the way for further research and improvements. An important aspect to stress is that this research has rarely reached outside the remit of NLP, and relatively few applications have involved humanities research (e.g., [39, 40, 161]). This is not particularly surprising, as it usually takes time for foundational research to find its way into application areas. However, as pointed out before (cf. [162]), given the high relevance of semantic change research for the analysis of concept evolution, this lack of disciplinary dialogue and exchange is a limiting factor and we hope that it will be addressed by future multidisciplinary research projects.

6. NLP for generating ontological structures

While automatic detection of lexical semantic change has shown advances in recent years despite a still insufficient interdisciplinary dialogue, the field of generating ontologies from diachronic corpora and representing them as linked data on the Web needs also further development of multidisciplinary approaches and exchanges, given the inherent complexity of the work involved. In this section, we discuss the main aspects pertaining to this type of task (block 4, Fig. 1), by taking account of previous research in areas such as ontology learning, construction of ontological diachronic structures from texts and automatic generation of linked data.

6.1. *Ontology learning*

Iyer et al. [163] survey the various approaches for (semi-)automatic ontology extraction and enrichment from unstructured text, including research papers from 1995 to 2018. They identify four broad categories of algorithms (similarity-based clustering, set-theoretic approach, Web corpus-based and deep learning) allowing for different types of ontology creation and updating, from clustering concepts in a hierarchy to learning and generating ontological representations for concepts, attributes and attribute restrictions. The authors perform an in-depth analysis of four "seminal algorithms" representative for each category (guided agglomerative clustering, C-PANKOW, formal concept analysis and word2vec) and compare them using ontology evaluation measures such as contextual relevance, precision and algorithmic efficiency. They also propose a deep learning method based on LSTMs, to tackle the problem of filtering out irrelevant data from corpora and improve relevance of retained concepts in a scalable manner.

Asim et al. [164] base their survey on the so-called "ontology learning layer cake" (introduced by Buitelaar et al. [165]), which illustrates the step-wise process of ontology acquisition starting with *terms*, and then moving up to *concepts, concept hierarchy, relations, relation hierarchy, axioms schemata*, and finally *axioms*. The paper categorises ontology learning techniques into linguistic, statistical and logical techniques, and presents detailed analysis and evaluation thereof. For instance, good performance is reported in the linguistic category for (lexico-)syntactic parsing and dependency analysis applied in relation extraction from texts in various domains and languages. C/NC-value (see also 6.3) and hierarchical clustering from the statistical group are featured for the tasks of acquiring concepts and relations respectively, while inductive logical programming from the logical group is mentioned for both tasks. Among the tools making use of such techniques considered by the authors as most prominent and widely used for ontology learning from text are Text2Onto [166], ASIUM [167] and CRCTOL [168], in the category hybrid (linguistic and statistical), OntoGain [169] and OntoLearn [170], solely based on statistical methods, and TextStorm/-Clouds [171] and Syndikate [172], from the logical category. Domain-specific or more wide-ranging datasets, such as Reuters-21578 [17] and British National Corpus [18], are also included in the description, as commonly used for testing and evaluating different ontology learning systems. Although published just one year earlier than [163], the survey does not mention any techniques based on neural networks. However, the authors state that ontology learning can benefit from incorporating deep learning methods into the field. Importantly, Asim et al. advocate for language independent ontology learning and for the necessity of human intervention in order to boost the overall quality of the outcome.

[17] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[18] http://www.natcorp.ox.ac.uk/

6.2. Diachronic constructs

He et al. [14] use the ontology learning layer cake framework and a diachronic corpus in Chinese (People's Daily Corpus), spanning from 1947 to 1996, to construct a set of diachronic ontologies by year and period. Their ontology learning system deals only with the first four bottom layers of the 'cake' (see also [164] and [165] above), for term extraction, synonymy recognition, concept discovery and hierarchical concept clustering. The first layer is built by segmenting and part of speech (POS) tagging the raw text using a hierarchical hidden Markov model (HHMM) for Chinese lexical analysis [173] and retaining all the words, except for stopwords and low frequency items. For synonymy detection, He et al. apply a distributional semantic model taking into account both lexical and syntactic contexts to compute the similarity between two terms, a method already utilised in diachronic corpus analysis in [174]. Cosine similarity and Kleinberg's "hubs and authorities" methodology [175] are used to group terms and synonyms into concepts and to select the top two terms with highest authority as semantic tags or labels for the concepts. An iterative K-means algorithm [176] is adopted to create a hierarchy of concepts with highly semantically associated clusters and sub-clusters. He et al. employ this four-step approach to build yearly/period diachronic XML ontologies for the considered corpus and evaluate concept discovery and clustering by comparing their results with a baseline computed via a Google word2vec implementation. The authors report that the proposed method outperformed the baseline in both concept discovery and hierarchical clustering, and that their diachronic ontologies were able to capture semantic changes of a term through comparison of its neighbouring terms or clusters at different points in time, and detect the apparition of new topics in a specific era. [14] also provides examples of diachronic analysis based on the ontologies derived from the studied corpus, such as shift in meaning from a domain to another, semantic change leading to polysemy or emergence of new similar terms as a result of real-world phenomena occurring in the period covered by the considered textual sources.

Other papers addressed the question of conceptualising semantic change using NLP techniques and diachronic corpora [144, 177, 178] implying various degrees of ontological formalisation.

Focusing on the way conceptual structures and the hierarchical relations among their components evolve over time, Bizzoni et al. [144] explore the direction of using hyperbolic embeddings for the construction of corpus-induced diachronic ontologies (see also Section 5.4). Using as a dataset the Royal Society Corpus, with a time span from 1665 to 1869, they show that such a method can detect symptoms of hierarchisation and specialisation in scientific language. Moreover, they argue that this type of technology may offer a (semi-)automatic alternative to the hand-crafted historical ontologies that require considerable amount of human expertise and skills to build hierarchies of concepts based on beliefs and knowledge of a different time.

In their analysis of changing relationships in temporal corpora, Rosin and Radinsky [177] propose several methods for constructing timelines that support the study of evolving languages. The authors introduce the task of timeline generation that implies two components, one for identifying "turning points", i.e. points in time when the target word underwent significant semantic changes, the other for identifying associated descriptors, i.e. words and events, that explain these changes in relation with real-world triggers. Their methodology includes techniques such as "peak detection" in time series and "projected embeddings", in order to define the timeline turning points and create a joint vector space for words and events, representing a specific time period. Different approaches are tested to compare vector representations of the same word or select the most relevant events causing semantic change over time, such as orthogonal Procrustes [10], similarity-based measures, and supervised machine learning (random forest, SVM and neural networks). After assessing these methods on datasets from Wikipedia, the New York Times archive and DBpedia, Rosin and Radinsky conclude that the best results are yielded by a supervised approach leveraging the projected embeddings, and the main factors affecting the quality of the created timelines are word ambiguity and the available amount of data and events related to the target word. Although [177] does not explicitly refer to ontology acquisition as a whole, automatic timeline generation provides insight into the modalities of detecting and conceptualising semantic change and word-event-time relationships that may serve with the task of corpus-based diachronic ontology generation.

Gulla et al. [178] make use of "concept signatures", representations constructed automatically from textual descriptions of existing concepts, to capture semantic changes of concepts over time. A concept signature is represented as a vector of weights. Each element in the vector corresponds to a linguistic unit or term (e.g. noun or noun phrase) extracted from the textual description of the concept, with its weight calculated as a tf-idf (term frequency - inverted document frequency) score. The process of signature building includes POS tagging, stopword removal, lemmatisation, noun/phrase selection and tf-idf computing for the selected linguistic units. According to Gulla et al., this type of vector representation enable comparisons via standard information retrieval measures, such as cosine similarity and Euclidian distance, that can uncover semantic drift of concepts in the ontology, both with respect to real-world phenomena (*extrinsic drift*) and inter-concept (taxonomic and non-taxonomic) relationships (*intrinsic drift*). The proposed methodology is applied to an ontology based on the Det Norske Veritas (DNV) company's Web site, [19] each Web page representing a concept. The text of the Web pages is used as a source for understanding the concepts and constructing the corresponding signatures at different points in time. [178] illustrates this procedure for various types of vector-based concept and relation comparison in the DNV ontology, computed for 2004 and 2008. The authors note that the size of the textual descriptions of concepts is determinant for the signature quality (too short descriptions may result in poor quality) and mention as further direction of research the use of deeper grammatical analysis of sentences and of semantic lexica for signature generation. Moreover, Gulla et al. point out that since the automatic construction of signatures relies on textual descriptions of existing concepts, the approach is primarily intended to updating existing structures rather than developing new ontologies.

6.3. Generating linked data

The transformation of the extracted information into formal descriptions that can be published as linked data on the Web is an important aspect of the process of ontology generation from textual sources. A number of tools have been devised to implement an integrated workflow for extracting concepts and relations, and converting the derived ontological structure into Semantic Web formalisations. While the first and second sub-sections above provided an overview of various approaches for corpus-based production of ontologies and ontological constructs including a temporal dimension, this sub-section focuses on means for making the generated output available on the Web in a structured and re-usable format. Three categories of tools dedicated to such tasks are discussed, for extracting information and linking entities to available ontologies on the Web, learning ontologies and translating the resulting models into Semantic Web representations, and for performing shallow conversion to RDF.

An example from the first category is LODifier [179], which combines different NLP techniques for named entity recognition, word sense disambiguation and semantic analysis to extract entities and relations from text and produce RDF representations linked to the LOD cloud using DBpedia and WordNet 3.0 vocabularies. The tool was evaluated on an English benchmark dataset containing newspapers, radio and television news from 1998.

From the second category, OntoGain [169] is a platform for unsupervised ontology acquisition from unstructured text. The concept identification module is based on C/NC-value [180], a method that enables the extraction of multi-word and nested terms from text. For the detection of taxonomic and non-taxonomic relations, [169] applies techniques such as agglomerative hierarchical clustering and formal concept analysis in the first task, and association rules and conditional probabilities in the second. OntoGain allows for the transformation of the resulted ontology into standard OWL statements. The authors report assessment including experiments with corpora from the medical and computer science domain, and comparisons with hand-crafted ontologies and similar applications such as Text2Onto.

Concept-Relation-Concept Tuple-based Ontology Learning (CRCTOL) [168] is a system for automatically mining ontologies from domain-specific documents. CRCTOL adopts various NLP methods such as POS tagging, multi-word extraction and tf-idf-based relevance measures for concept learning, a variant of Lesk's algorithm [181] for word sense disambiguation, and WordNet hierarchy processing and full text parsing for the construction of taxonomic and non-taxonomic relations. The derived ontology is then modelled as a graph, with the possibility of exporting the corresponding representation in RDFS and OWL format. [168] presents two case studies, for building a terrorism domain ontology and a sport event domain ontology, as well as results of quantitative and qualitative evaluation of the tool through various comparisons with other systems or assessment references such as

[19] A company specialising in risk management and certification.

Text-To-Onto/Text2Onto, WordNet, expert rating and human-edited benchmark ontologies.

One of the systems often cited as a reference in ontology learning from textual resources (see also above) is Text2Onto (the successor of TextToOnto) [166]. Based on the GATE framework, it combines linguistic pre-processing (e.g. tokenisation, sentence splitting, POS tagging, lemmatisation) with the use of a JAPE transducer and shallow parsing run on the pre-processed corpus to identify concepts, instances and different types of relations (subclass-of, part-of, instance-of, etc.) to be included in a Probabilistic Ontology Model (POM). The model, independent of any knowledge representation formalism, can be then translated into various ontology representation languages such as RDFS, OWL and F-Logic. The paper also describes a strategy for data-driven change discovery allowing for selective POM updating and traceability of the ontology evolution, consistent with the changes in the underlying corpus. Evaluation is reported with respect to certain tasks and a collection of tourism-related texts, the results being compared with a reference taxonomy for the domain.

Recent work accounts for more specialised tools, from the third category, such as converters, making, for instance, linked data in RDF format out of CSV files (CoW [20] and cattle [21] [4]) or directly converting language resources into LL(O)D (LLODifier [22] [182]). As already pointed out at the beginning of this section, the field may benefit from further exchanges among scholars in different areas of studies such as theoretical and cognitive linguistics, history and philosophy of language, digital humanities, NLP and Semantic Web.

7. LL(O)D resources and publication

In this section (related to block 5, Fig. 1), we outline the existing resources on the Web including diachronic representation of data from the humanities, with a view towards the possibilities of integrating more resources of this kind into the LL(O)D cloud in the future.

The main nucleus for linguistic linked open data is the LL(O)D cloud [183], [23] which started in 2011 with less than 30 datasets, and at the time of writing consists of 185 different datasets. The resources linked in the LL(O)D cloud include corpora, lexicons and dictionaries, terminologies, thesauri and knowledge bases, linguistic resources metadata, linguistic data categories, and typological databases. The LL(O)D diagram is generated automatically from the subset of Linghub [24] that is published as linked open data.

Not all diachronic datasets are registered through Linghub/LL(O)D Cloud. Within the CLARIAH project [25] several datasets have been converted from CSV format to linked open data, and published through project websites or GitHub. For example, in [184], different diachronic lexicons are modelled according to the Lemon model and interlinked, such that one can query across time and dialect variations.

Also in the Netherlands, the Amsterdam Time Machine connects attestations of Amsterdam dialects and sociolects, cinema and theatre locations and tax information to base maps of Amsterdam at various points in time [185]. A combined resource like this, allows scholars to investigate ‘higher’ and ‘lower’ sociolects in conjunction with ‘elite density’ in a neighbourhood (i.e. the proportion of wealthier people that lived in an area). Lexicologists at the Dutch Language Institute have been creating dictionaries of Dutch that cover the period from 500 to 1976 which are now being modelled through OntoLex-Lemon [186].

Searching for and modelling diachronic change requires rethinking some contemporary (semantic) Web infrastructure. As [187] shows, standardised language tags cannot capture the differences between Old-, Middle- and Modern French resources.

Digital editions, often modelled in TEI [188], are a rich resource of diachronic language variation. Some corpora, such as the 15th-19th-century Spanish poetry corpus described in [189] contain additional annotations such as psychological and affective labels, but it seems the study was not focused particularly on how these aspects may have changed over time.

For humanities scholars such as historians, who deal with source materials dating back to for example the early modern period, language change is a given, but the knowledge they gain over time is not always formalised or published as linked data. For example, a project that analyses the representation of emotions plays from the 17th to the 19th century, a dataset and lexicon were developed, but these were not explicitly

[20]https://pypi.org/project/cow-csvw/
[21]http://cattle.datalegend.net/
[22]https://github.com/acoli-repo/LLODifier
[23]https://linguistic-lod.org/
[24]http://linghub.org
[25]https://clariah.nl

linked to the (linguistic) LL(O)D cloud [190, 191]. [26] In contrast to [189], here the labels are explicitly grounded in time. There is a task here for the Semantic Web community to make it easier to publish and maintain LL(O)D datasets for non-Semantic Web experts.

It should be also noted that while there do not currently exist guidelines for publishing lexicons and ontologies representing semantic change as LL(O)D data, there are moves towards producing such material within the *Nexus Linguarum* COST Action, however, with particular reference to the overlap between different working groups and UC4.2.1.

8. Conclusions

This paper presents a literature survey, bringing together various fields of research that may be of interest in the construction of a workflow for detecting and representing semantic change (Fig. 1). The state of the art described in the paper also represents the starting point in designing a methodology, based on this workflow, for the humanities use case UC4.2.1 as an application within the COST Action *Nexus Linguarum, European network for Web-centred linguistic data science*. The survey touches upon the use of multilingual diachronic corpora from the humanities, and different approaches from linguistics-related disciplines, NLP and Semantic Web. The organisation of the sections and the themes included in the outline reflects the heterogeneity and complexity of the task and the necessity of a framework enabling interdisciplinary dialogue and collaboration.

At this stage, the reviewed literature and main surveyed approaches and tools (see Appendix) suggest that the theoretical frameworks (Section 3) and the NLP techniques for detecting lexical semantic change (Section 5) show good levels of development, although certain conceptual and technical difficulties are yet to overcome. The fields dealing with the generation of diachronic ontologies from unstructured text and their representation as LL(O)D formalisms on the Web (Section 4, 6, 7) would require further harmonisation with the previous points and research investment.

Despite recent advances in creating and publishing linguistic resources on the LL(O)D cloud, and the availability of potentially relevant resources, humanities researchers working on the detection and representation of semantic change as linked data on the Web are still confronted with a series of challenges. These include limitations in representing temporal and dynamic aspects given the work in progress status of some of the applicable semantic Web technologies, absence of guidelines for producing diachronic ontologies, and lack of ways to ease publication and maintenance of data for non-Semantic Web experts. Another point requiring further attention is the need for building connections between the various areas of research involved in the type of task described in the paper. As we tried to illustrate through the structure of the generic workflow and the discussions within the related sections, the research agenda for attaining this goal should include interdisciplinary approaches and exchanges among the identified fields of study. The results of the survey seem to suggest that there are not yet enough interrelations and explicit connections between these fields, and the area under investigation would benefit from further developments in this direction.

We assume that, given the current progress in deep learning, digital humanities and the ongoing undertakings in LL(O)D, the detection and representation of semantic change as linked data combined with the analysis of large datasets from the humanities will acquire the level of attention and dialogue needed for the advancement in this area of study. Detecting and representing semantic change as LL(O)D is an important topic for the future development of Semantic Web technologies, since learning to deal with the knowledge of the past and its evolution over time also implies learning to deal with the knowledge of the future.

[26] https://www.esciencecenter.nl/projects/from-sentiment-mining-to-mining-embodied-emotions/

References

[1] T. Burrows, E. Hyvönen, L. Ransom and H. Wijsman, Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts, *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies* **3**(1) (2018), 249–252.

[2] L. Isaksen, R. Simon, E.T. Barker and P. de Soto Cañamares, Pelagios and the emerging graph of ancient world data, in: *Proceedings of the 2014 ACM conference on Web science*, 2014, pp. 197–201.

[3] A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach and F. van Harmelen, Semantic technologies for historical research: A survey, *Semantic Web* **6**(6) (2014), 539–564–. doi:10.3233/SW-140158.

[4] A. Meroño-Peñuela, V. de Boer, M. van Erp, W. Melder, R. Mourits, R. Schalk and R. Zijdeman, Ontologies in CLAR-

IAH: Towards Interoperability in History, Language and Media, *https://arxiv.org/abs/2004.02845v2* (2020), 26.

[5] C. Chiarcos and A. Pareja-Lora, *Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction*, in: *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*, A. Pareja-Lora, M. Blume, B.C. Lust and C. Chiarcos, eds, MIT Press, 2019, pp. 1–18. ISBN 978-0-262-53625-7.

[6] N. Tahmasebi, L. Borin and A. Jatowt, Survey of Computational Approaches to Lexical Semantic Change, *arXiv: Computation and Language* (2018).

[7] X. Tang, A state-of-the-art of semantic change computation, *Natural Language Engineering* **24**(5) (2018), 649–676–. doi:10.1017/S1351324918000220.

[8] A. Kutuzov, L. Øvrelid, T. Szymanski and E. Velldal, Diachronic word embeddings and semantic shifts: A survey, in: *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1384–1397.

[9] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky and N. Tahmasebi, SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection, in: *Proceedings of the 14th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Barcelona, Spain, 2020.

[10] W.L. Hamilton, J. Leskovec and D. Jurafsky, Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 2016, pp. 1489–1501.

[11] S. Wang, S. Schlobach and M. Klein, Concept drift and how to identify it, *Journal of Web Semantics First Look* (2011). http://dx.doi.org/10.2139/ssrn.3199520.

[12] A. Fokkens, S. Ter Braake, I. Maks and D. Ceolin, On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change, *Drift-a-LOD@EKAW* (2016).

[13] T. McEnery and A. Hardie, *Corpus-based studies of synchronic and diachronic variation*, in: *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press, 2011, pp. 94–121–. ISBN 978-0-511-98139-5. doi:10.1017/CBO9780511981395. http://ebooks.cambridge.org/ref/id/CBO9780511981395.

[14] S. He, X. Zou, L. Xiao and J. Hu, Construction of Diachronic Ontologies from People's Daily of Fifty Years, *LREC 2014 Proceedings* (2014).

[15] M. Richter, *The History of Political and Social Concepts: A Critical Introduction*, Oxford University Press, 1995.

[16] J.-M. Kuukkanen, Making Sense of Conceptual Change **47**(3) (2008), 351–372. doi:https://doi.org/10.1111/j.1468-2303.2008.00459.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2303.2008.00459.x.

[17] T.G. Stavropoulos, S. Andreadis, M. Riga, E. Kontopoulos, P. Mitzias and I. Kompatsiaris, A Framework for Measuring Semantic Drift in Ontologies, 2016.

[18] M. Fitting, Intensional Logic, in: *The Stanford Encyclopedia of Philosophy*, Spring 2020 edn, E.N. Zalta, ed., Metaphysics Research Lab, Stanford University, 2020. https://plato.stanford.edu/archives/spr2020/entries/logic-intensional/.

[19] F. de Saussure, *Cours de linguistique générale (1916)*, Payot, 1971. https://fr.wikisource.org/wiki/Cours_de_linguistique_générale/Texte_entier.

[20] A. Betti and H. van den Berg, Modelling the History of Ideas, *British Journal for the History of Philosophy* **22**(4) (2014), 812–835. doi:10.1080/09608788.2014.949217.

[21] G. Widmer and M. Kubat, Learning in the presence of concept drift and hidden contexts, *Machine Learning, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands* **23**(1) (1996), 69–101–.

[22] G. Antoniou, M. d'Aquin and J.Z. Pan, Semantic Web dynamics, *Journal of Web Semantics* **9**(3) (2011), 245–246–. doi:10.1016/j.websem.2011.06.008.

[23] N.B. Kvastad, Semantics in the Methodology of the History of Ideas, *Journal of the History of Ideas, University of Pennsylvania Press* **38**(1) (1977), 157–174–.

[24] D. Geeraerts, *Theories of lexical semantics*, Oxford University Press, 2010. ISBN 978-0-19-870031-9.

[25] S. Grondelaers, D. Speelman and D. Geeraerts, Lexical variation and change, in: *The Oxford handbook of cognitive linguistics*, 2007.

[26] C. Roche, Ontoterminology: How to unify terminology and ontology into a single paradigm, in: *LREC 2012 - Eighth international conference on Language Resources and Evaluation*, 2012, pp. 2626–2630–. http://christophe-roche.fr/Bibliographie/2012/567_Paper_Header.pdf.

[27] D. Gromann, Terminology meets the multilingual Semantic Web: A semiotic comparison of ontologies and terminologies, in: *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes*, G. Budin and V. Lušicky, eds, University of Vienna, 2013, pp. 418–428–. ISBN 978-3-200-03674-1.

[28] R. Temmerman, *Towards New Ways of Terminology Description: The sociocognitive approach*, Terminology and Lexicography Research and Practice, Vol. 3, John Benjamins Publishing Company, 2000. ISBN 978-90-272-2326-5. doi:10.1075/tlrp.3. http://www.jbe-platform.com/content/books/9789027298638.

[29] D. Schiffrin, *Discourse markers*, Vol. 5, Cambridge University Press, 1987.

[30] B. Fraser, Pragmatic markers, *Pragmatics* **6**(2) (1996), 167–190.

[31] K. Aijmer, I think–an English modal particle, *Modality in Germanic languages: Historical and comparative perspectives* **1** (1997), 47.

[32] B. Fraser, What are discourse markers?, *Journal of pragmatics* **31**(7) (1999), 931–952.

[33] D. Schiffrin, Discourse marker research and theory: revisiting and, *Approaches to discourse particles* **1** (2006), 315–338.

[34] P. Auer and Y. Maschler, *NU/NÅ: A family of discourse markers across the languages of Europe and beyond*, Vol. 58, Walter de Gruyter GmbH & Co KG, 2016.

[35] R. Waltereit and U. Detges, Different functions, different histories. Modal particles and discourse markers from a diachronic point of view, *Catalan journal of linguistics* (2007), 61–80.

[36] L.S. Stvan, Diachronic change in the uses of the discourse markers why and say in American English, *Linguistic Insights-Studies in Language and Communication* **25** (2006), 61–76.

[37] L. Downing, The Cambridge Introduction to Michel Foucault (2008).

[38] R. Wodak, *Critical Discourse Analysis, Discourse-Historical Approach*, in: *The International Encyclopedia of Language and Social Interaction*, 1st edn, K. Tracy, T. Sandel and C. Ilie, eds, Wiley, 2015. ISBN 978-1-118-61110-4. doi:10.1002/9781118611463.

[39] L. Viola and J. Verheul, One Hundred Years of Migration Discourse in The Times: A Discourse-Historical Word Vector Space Approach to the Construction of Meaning, *Frontiers in Artificial Intelligence* **3** (2020), 64. doi:10.3389/frai.2020.00064.

[40] S. Soni, L. Klein and J. Eisenstein, Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers, *arXiv:2103.07538 [cs]* (2021), arXiv: 2103.07538. http://arxiv.org/abs/2103.07538.

[41] J.P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar and P. Cimiano, The OntoLex-Lemon Model: Development and Applications (2017), 587–597, Publisher: Lexical Computing CZ s.r.o. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

[42] N. Guarino, D. Oberle and S. Staab, What Is an Ontology?, in: *Handbook on Ontologies*, S. Staab and R. Studer, eds, International Handbooks on Information Systems, Springer, Berlin, Heidelberg, 2009, pp. 1–17. ISBN 9783540926733. doi:10.1007/978-3-540-92673-3_0. https://doi.org/10.1007/978-3-540-92673-3_0.

[43] C. Chiarcos, M. Ionov, J. de Does, K. Depuydt, A.F. Khan, S. Stolk, T. Declerck and J.P. McCrae, Modelling Frequency and Attestations for OntoLex-Lemon, in: *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, European Language Resources Association, Marseille, France, 2020, pp. 1–9. ISBN 979-10-95546-46-7. https://www.aclweb.org/anthology/2020.globalex-1.1.

[44] S. Salmon-Alt, Data structures for etymology: towards an etymological lexical network., *BULAG* **31** (2006), 1–12.

[45] J. Bowers and L. Romary, Deep encoding of etymological information in TEI, *Journal of the Text Encoding Initiative* (2016).

[46] L. Romary, M. Khemakhem, F. Khan, J. Bowers, N. Calzolari, M. George, M. Pet and P. Bański, LMF Reloaded, *arXiv preprint arXiv:1906.02136* (2019).

[47] F. Khan, L. Romary, A. Salgado, J. Bowers, M. Khemakhen and T. Tasovac, Modelling Etymology in LMF/TEI, in: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), 2020.

[48] G. de Melo, Etymological Wordnet: Tracing The History of Words., in: *Proceedings of the 9th Conference on Language Resources and Evaluation (LREC 2014)*, European Language Resources Association (ELRA), 2014.

[49] C. Chiarcos, F. Abromeit, C. Fäth and M. Ionov, Etymology Meets Linked Data. A Case Study In Turkic., in: *Digital Humanities 2016. Krakow*, 2016.

[50] F. Khan, Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA). eventplace: Miyazaki, Japan*, 2018.

[51] F. Khan, J. Díaz-Vera and M. Monachini, Representing meaning change in computational lexical resources; the case of shame and embarrassment in Old English, *Formal Representation and the Digital Humanities* (2018), 59.

[52] F. Mambrini and M. Passarotti, Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin, in: *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, European Language Resources Association, Marseille, France, 2020, pp. 20–28. ISBN 979-10-95546-46-7. https://www.aclweb.org/anthology/2020.globalex-1.3.

[53] F. Khan, A. Bellandi and M. Monachini, Tools and Instruments for Building and Querying Diachronic Computational Lexica, in: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, The COLING 2016 Organizing Committee, pp. 164–171. https://www.aclweb.org/anthology/W16-4022.

[54] F. Rizzolo, Y. Velegrakis, J. Mylopoulos and S. Bykau, Modeling concept evolution: a historical perspective, in: *International Conference on Conceptual Modeling*, Springer, 2009, pp. 331–345.

[55] C. Gutierrez, C. Hurtado and A. Vaisman, Temporal RDF, in: *The Semantic Web: Research and Applications*, A. Gómez-Pérez and J. Euzenat, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 93–107. ISBN 978-3-540-31547-6.

[56] P. Garbacz and R. Trypuz, Representation of Tensed Relations in OWL, in: *Metadata and Semantic Research*, Vol. 755, E. Garoufallou, S. Virkus, R. Siatri and D. Koutsomiha, eds, Springer International Publishing, 2017, pp. 62–73, Series Title: Communications in Computer and Information Science. ISBN 978-3-319-70862-1 978-3-319-70863-8. doi:10.1007/978-3-319-70863-8_6. http://link.springer.com/10.1007/978-3-319-70863-8_6.

[57] C. Welty, R. Fikes and S. Makarios, A reusable ontology for fluents in OWL, in: *FOIS*, Vol. 150, 2006, pp. 226–236.

[58] H.-U. Krieger, A detailed comparison of seven approaches for the annotation of time-dependent factual knowledge in RDF and OWL, in: *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 2014, p. 1.

[59] F. Khan and J. Bowers, Towards a Lexical Standard for the Representation of Etymological Data, in: *Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale*, 2020.

[60] J.R. Hobbs and F. Pan, Time ontology in OWL, *W3C working draft* **27** (2006), 133.

[61] J.F. Allen, Maintaining knowledge about temporal intervals, *Communications of the ACM* **26**(11) (1983), 832–843.

[62] S. Batsakis, E.G. Petrakis, I. Tachmazidis and G. Antoniou, Temporal representation and reasoning in OWL 2, *Semantic Web* **8**(6) (2017), 981–1000.

[63] P. Golden and R. Shaw, Period assertion as nanopublication: The PeriodO period gazetteer, in: *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1013–1018.

[64] P. Cook, J.H. Lau, D. McCarthy and T. Baldwin, Novel word-sense identification, in: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1624–1635.

[65] J.H. Lau, P. Cook, D. McCarthy, S. Gella and T. Baldwin, Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models, in: *Proceedings of the 52nd Annual Meeting of the Associa-*

tion for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2014, pp. 259–270.

[66] L. Fremann and M. Lapata, A Bayesian model of diachronic meaning change, *Transactions of the Association for Computational Linguistics* **4** (2016), 31–45.

[67] S. Mitra, R. Mitra, S.K. Maity, M. Riedl, C. Biemann, P. Goyal and A. Mukherjee, An automatic approach to identify word sense changes in text media across timescales, *Natural Language Engineering* **21**(5) (2015), 773–798.

[68] N. Tahmasebi and T. Risse, Finding Individual Word Sense Changes and their Delay in Appearance, in: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 2017, pp. 741–749.

[69] Y. Kim, Y. Chiu, K. Hanaki, D. Hegde and S. Petrov, Temporal Analysis of Language through Neural Language Models, in: *LTCSS@ACL*, Association for Computational Linguistics, 2014, pp. 61–65.

[70] P. Basile and B. McGillivray, *Discovery Science*, in *Lecture Notes in Computer Science*, Vol. 11198, Springer-Verlag, 2018, Chapter Exploiting the Web for Semantic Change Detection.

[71] V. Kulkarni, R. Al-Rfou, B. Perozzi and S. Skiena, Statistically significant detection of linguistic change, in: *Proceedings of the 24th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2015, pp. 625–635.

[72] H. Dubossarsky, D. Weinshall and E. Grossman, Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1136–1145.

[73] N. Tahmasebi, A Study on Word2Vec on a Historical Swedish Newspaper Corpus, in: *CEUR Workshop Proceedings. Vol. 2084. Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, Helsinki Finland, March 7-9, 2018.*, University of Helsinki, Faculty of Arts, Helsinki, 2018.

[74] M. Rudolph and D. Blei, Dynamic Embeddings for Language Evolution, in: *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018, pp. 1003–1011.

[75] A. Jatowt, R. Campos, S.S. Bhowmick, N. Tahmasebi and A. Doucet, Every Word has its History: Interactive Exploration and Visualization of Word Sense Evolution, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ACM, 2018, pp. 1899–1902.

[76] A. Kutuzov, Distributional word embeddings in modeling diachronic semantic change, PhD thesis, University of Oslo, 2020.

[77] V. Perrone, M. Palma, S. Hengchen, A. Vatri, J.Q. Smith and B. McGillivray, GASC: Genre-Aware Semantic Change for Ancient Greek, in: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 56–66. https://www.aclweb.org/anthology/W19-4707.

[78] H. Dubossarsky, S. Hengchen, N. Tahmasebi and D. Schlechtweg, Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Florence, Italy, 2019.

[79] P. Shoemark, F. Ferdousi Liza, D. Nguyen, S. Hale and B. McGillivray, Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, 2019, pp. 66–76.

[80] M. Piotrowski, *Natural Language Processing for Historical Texts*, Morgan & Claypool, 2012.

[81] B. McGillivray, *Methods in Latin Computational Linguistics*, Brill, Leiden, 2014.

[82] P. Rayson, D.E. Archer, A. Baron, J. Culpeper and N. Smith, Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora, in: *Proceedings of the Corpus Linguistics conference: CL2007*, 2007.

[83] S. Scheible, R.J. Whitt, M. Durrell and P. Bennett, Evaluating an 'off-the-shelf'POS-tagger on Early Modern German text, in: *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, 2011, pp. 19–23.

[84] M. Bollmann, A large-scale comparison of historical text normalization systems, *arXiv preprint arXiv:1904.02036* (2019).

[85] A. Baron and P. Rayson, VARD2: A tool for dealing with spelling variation in historical corpora, in: *Postgraduate conference in corpus linguistics*, 2008.

[86] M. Bollmann, automatic normalization of historical texts using distance measures and the Norma tool, in: *Proceedings of the second workshop on annotation of corpora for research in the humanities (ACRH-2), Lisbon, Portugal*, 2012, pp. 3–14.

[87] M. Bollmann, F. Petran and S. Dipper, Rule-based normalization of historical texts, in: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, 2011, pp. 34–42.

[88] J. Porta, J.-L. Sancho and J. Gómez, Edit transducers for spelling variation in Old Spanish, in: *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, Linköping University Electronic Press, 2013, pp. 70–79.

[89] I. Etxeberria, I. Alegria, L. Uria and M. Hulden, Evaluating the noisy channel model for the normalization of historical texts: Basque, Spanish and Slovene, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 1064–1069.

[90] K. Jassem, F. Graliński, T. Obrębski and P. Wierzchoń, Automatic Diachronic Normalization of Polish Texts, *Investigationes Linguisticae* **37** (2017), 17–33.

[91] M. Kestemont, W. Daelemans and G. De Pauw, Weigh your words—memory-based lemmatization for Middle Dutch, *Literary and Linguistic Computing* **25**(3) (2010), 287–301.

[92] E. Pettersson, B. Megyesi and J. Nivre, Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting, in: *Proceedings of the 19th Nordic conference of computational linguistics (Nodalida 2013)*, 2013, pp. 163–179.

[93] Y. Adesam, M. Ahlberg and G. Bouma, bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa... Towards lexical link-up for a corpus of Old Swedish., in: *KONVENS*, 2012, pp. 365–369.

[94] H. van Halteren and M. Rem, Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters, *Language resources and evaluation* **47**(4) (2013), 1233–1259.

[95] C. Oravecz, B. Sass and E. Simon, Semi-automatic normalization of Old Hungarian codices, in: *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, 2010, pp. 55–59.

[96] F. Sánchez-Martínez, I. Martínez-Sempere, X. Ivars-Ribes and R.C. Carrasco, An open diachronic corpus of historical Spanish: annotation criteria and automatic modernisation of spelling, *arXiv preprint arXiv:1306.3692* (2013).

[97] E. Pettersson, Spelling normalisation and linguistic analysis of historical text for information extraction, PhD thesis, Acta Universitatis Upsaliensis, 2016.

[98] M. Domingo and F. Casacuberta, Spelling normalization of historical documents by using a machine translation approach (2018).

[99] M. Bollmann, J. Bingel and A. Søgaard, Learning attention for historical text normalization by learning to pronounce, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 332–344.

[100] N. Korchagina, Normalizing Medieval German Texts: from rules to deep learning, in: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 2017, pp. 12–17.

[101] A. Robertson and S. Goldwater, Evaluating Historical Text Normalization Systems: How Well Do They Generalize?, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 720–725.

[102] M. Hämäläinen, T. Säily, J. Rueter, J. Tiedemann and E. Mäkelä, Normalizing early English letters to present-day English spelling, in: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2018, pp. 87–96.

[103] S. Flachs, M. Bollmann and A. Søgaard, Historical Text Normalization with Delayed Rewards, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1614–1619.

[104] M.A. Azawi, M.Z. Afzal and T.M. Breuel, Normalizing historical orthography for OCR historical documents using LSTM, in: *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, 2013, pp. 80–85.

[105] M. Bollmann and A. Søgaard, Improving historical spelling normalization with bi-directional LSTMs and multi-task learning, *arXiv preprint arXiv:1610.07844* (2016).

[106] M. Kestemont, G. De Pauw, R. van Nie and W. Daelemans, Lemmatization for variation-rich languages using deep learning, *Digital Scholarship in the Humanities* **32**(4) (2017), 797–815.

[107] P. Mitankin, S. Gerdjikov and S. Mihov, An approach to unsupervised historical text normalisation, in: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 29–34.

[108] N. Ljubešic, K. Zupan, D. Fišer and T. Erjavec, Normalising Slovene data: historical texts vs. user-generated content, in: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Vol. 16, 2016, pp. 146–155.

[109] S. Soni, K. Lerman and J. Eisenstein, Follow the Leader: Documents on the Leading Edge of Semantic Change Get More Citations, *arXiv:1909.04189 [physics]* (2020), arXiv: 1909.04189. http://arxiv.org/abs/1909.04189.

[110] S. Montariol, M. Martinc and L. Pivovarova, Scalable and Interpretable Semantic Change Detection, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 4642–4652–. doi:10.18653/v1/2021.naacl-main.369. https://www.aclweb.org/anthology/2021.naacl-main.369.

[111] D. van Strien, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray and G. Colavizza, Assessing the Impact of OCR Quality on Downstream NLP Tasks:, in: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, SCITEPRESS - Science and Technology Publications, 2020, pp. 484–496–. ISBN 978-989-758-395-7. doi:10.5220/0009169004840496.

[112] L. Borin, D. Kokkinakis and L.-J. Olsson, Naming the past: Named entity and animacy recognition in 19th century Swedish literature, in: *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, 2007, pp. 1–8.

[113] C. Grover, S. Givon, R. Tobin and J. Ball, Named Entity Recognition for Digitised Historical Texts, in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[114] K. Kettunen and T. Ruokolainen, Names, right or wrong: Named entities in an OCRed historical Finnish newspaper collection, in: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, 2017, pp. 181–186.

[115] N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann and T. Risse, Neer: An unsupervised method for named entity evolution recognition, in: *Proceedings of COLING 2012*, 2012, pp. 2553–2568.

[116] C. Neudecker, L. Wilms, W.J. Faber and T. van Veen, Large-scale refinement of digital historic newspapers with named entity recognition, in: *Proc IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting*, 2014.

[117] S. Mac Kim and S. Cassidy, Finding names in trove: named entity recognition for Australian historical newspapers, in: *Proceedings of the Australasian Language Technology Association Workshop 2015*, 2015, pp. 57–65.

[118] S.T. Aguilar, X. Tannier and P. Chastang, Named entity recognition applied on a data base of Medieval Latin charters. The case of chartae burgundiae, in: *3rd International Workshop on Computational History (HistoInformatics 2016)*, 2016.

[119] R. Sprugnoli, Arretium or Arezzo? a neural approach to the identification of place names in historical texts, in: *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, aAccademia University Press, 2018, pp. 360–365.

[120] M. Riedl and S. Padó, A named entity recognition shootout for german, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 120–125.

[121] H. Hubková, Named-entity recognition in Czech historical texts: Using a CNN-BiLSTM neural network model, 2019.

[122] K. Labusch, P. Kulturbesitz, C. Neudecker and D. Zellhöfer, BERT for Named Entity Recognition in Contemporary and Historical German, in: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 2019.

[123] S. Schweter and J. Baiter, Towards robust named entity recognition for historic german, *arXiv preprint arXiv:1906.07592* (2019).

[124] P. Agarwal, J. Strötgen, L. Del Corro, J. Hoffart and G. Weikum, Dianed: time-aware named entity disambiguation for diachronic corpora, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 686–693.

[125] M. Ehrmann, M. Romanello, A. Flückiger and S. Clematide, Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers, in: *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, Vol. 2696, CEUR, 2020.

[126] M. Ehrmann, M. Romanello, A. Flückiger and S. Clematide, Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2020, pp. 288–310.

[127] M. Rovera, F. Nanni, S.P. Ponzetto and A. Goy, Domain-specific named entity disambiguation in historical memoirs, in: *CEUR Workshop Proceedings*, Vol. 2006, RWTH, 2017, p. Paper–20.

[128] F. Frontini, C. Brando and J.-G. Ganascia, Semantic web based named entity linking for digital humanities and heritage texts, 2015.

[129] S. Van Hooland, M. De Wilde, R. Verborgh, T. Steiner and R. Van de Walle, Exploring entity recognition and disambiguation for cultural heritage collections, *Digital Scholarship in the Humanities* **30**(2) (2015), 262–279.

[130] M. De Wilde, S. Hengchen et al., Semantic enrichment of a multilingual archive with linked open data, *Digital Humanities Quarterly* (2017).

[131] C. Brando, F. Frontini and J.-G. Ganascia, REDEN: named entity linking in digital literary editions using linked data sets, *Complex Systems Informatics and Modeling Quarterly* (2016), 60–80.

[132] S. Rosset, C. Grouin, K. Fort, O. Galibert, J. Kahn and P. Zweigenbaum, Structured named entities in two distinct press corpora: contemporary broadcast news and old newspapers, in: *Proceedings of the Sixth Linguistic Annotation Workshop*, 2012, pp. 40–48.

[133] E.L. Pontes, L.A. Cabrera-Diego, J.G. Moreno, E. Boros, A. Hamdi, N. Sidère, M. Coustaty and A. Doucet, Entity Linking for Historical Documents: Challenges and Solutions, in: *International Conference on Asian Digital Libraries*, Springer, 2020, pp. 215–231.

[134] S. Rijhwani and D. Preoţiuc-Pietro, Temporally-informed analysis of named entity recognition, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7605–7617.

[135] K. Gulordava and M. Baroni, A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus., in: *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, 2011, pp. 67–71.

[136] C. Liebeskind, I. Dagan and J. Schler, Statistical thesaurus construction for a morphologically rich language, in: ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 59–64.

[137] A. Jatowt and K. Duh, A framework for analyzing semantic change of words across time, in: *IEEE/ACM Joint Conference on Digital Libraries*, IEEE, 2014, pp. 229–238.

[138] E. Sagi, S. Kaufmann and B. Clark, Tracing semantic change with latent semantic analysis, *Current methods in historical semantics* **73** (2011), 161–183.

[139] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, in: *International Conference on Learning Representations*, 2013, pp. 1–12.

[140] J. Pennington, R. Socher and C. Manning, GloVe: Global Vectors for Word Representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.

[141] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics* **5** (2017), 135–146. doi:10.1162/tacl_a_00051.

[142] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, Advances in Pre-Training Distributed Word Representations, in: *International Conference on Language Resources and Evaluation*, 2018, pp. 52–55.

[143] H. Gong, S. Bhat and P. Viswanath, Enriching Word Embeddings with Temporal and Spatial Information, in: *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, 2020, pp. 1–11. https://www.aclweb.org/anthology/2020.conll-1.1.

[144] Y. Bizzoni, M. Mosbach, D. Klakow and S. Degaetano-Ortlieb, Some steps towards the generation of diachronic WordNets, in: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 2019, pp. 55–64. https://www.aclweb.org/anthology/W19-6106.

[145] M. Nickel and D. Kiela, Poincaré Embeddings for Learning Hierarchical Representations, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6341–6350–.

[146] A. Tsakalidis and M. Liakata, Sequential Modelling of the Evolution of Word Representations for Semantic Change Detection, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 8485–8497. doi:10.18653/v1/2020.emnlp-main.682.

[147] A. Wegmann, F. Lemmerich and M. Strohmaier, Detecting Different Forms of Semantic Shift in Word Embeddings via Paradigmatic and Syntagmatic Association Changes, in: *Lecture Notes in Computer Science*, Springer International Pub-

lishing, 2020, pp. 619–635. doi:10.1007/978-3-030-62419-4_35.

[148] W.L. Hamilton, J. Leskovec and D. Jurafsky, Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2016, pp. 2116–2121. doi:10.18653/v1/D16-1229.

[149] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[150] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep Contextualized Word Representations, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. doi:10.18653/v1/N18-1202. https://www.aclweb.org/anthology/N18-1202.

[151] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q.V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, in: *Advances in Neural Information Processing Systems*, 2019, pp. 5753–5763.

[152] M. Giulianelli, M.D. Tredici and R. Fernández, Analysing Lexical Semantic Change with Contextualised Word Representations, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 3960–3973. doi:10.18653/v1/2020.acl-main.365.

[153] V. Kanjirangat, S. Mitrovic, A. Antonucci and F. Rinaldi, SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces, in: *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May and E. Shutova, eds, International Committee for Computational Linguistics, 2020, pp. 214–221. https://www.aclweb.org/anthology/2020.semeval-1.26/.

[154] D.M. Blei, A.Y. Ng and M.I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research* **3** (2003), 993–1022–.

[155] D.M. Blei and J.D. Lafferty, Dynamic topic models, in: *Proceedings of the 23rd international conference on Machine learning - ICML '06*, ACM Press, 2006, pp. 113–120–. ISBN 978-1-59593-383-6. doi:10.1145/1143844.1143859. http://portal.acm.org/citation.cfm?doid=1143844.1143859.

[156] C. Pölitz, T. Bartz, K. Morik and A. Störrer, *Investigation of Word Senses over Time Using Linguistic Corpora*, in: *Text, Speech, and Dialogue*, P. Král and V. Matoušek, eds, Lecture Notes in Computer Science, Vol. 9302, Springer International Publishing, 2015, pp. 191–198–. ISBN 978-3-319-24032-9. doi:10.1007/978-3-319-24033-6_22. http://link.springer.com/10.1007/978-3-319-24033-6_22.

[157] X. Wang and A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, in: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, ACM Press, 2006, p. 424. ISBN 978-1-59593-339-3. doi:10.1145/1150402.1150450. http://portal.acm.org/citation.cfm?doid=1150402.1150450.

[158] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei, Hierarchical Dirichlet Processes, *Journal of the American Statistical Association* **101**(476) (2006), 1566–1581–. doi:10.1198/016214506000000302.

[159] B. McGillivray, R. Buning and S. Hengchen, Topic Modelling: Hartlib's Correspondence before and after 1650, in: *Reassembling the Republic of Letters in the Digital Age*, H. Hotson and T. Wallnig, eds, Göttingen University Press, 2019.

[160] S. Hengchen, When does it mean? Detecting semantic change in historical texts, PhD thesis, Université libre de Bruxelles, 2017.

[161] B. McGillivray, S. Hengchen, V. Lähteenoja, M. Palma and A. Vatri, A computational approach to lexical polysemy in Ancient Greek, *Digital Scholarship in the Humanities* **34**(4) (2019), 893–907.

[162] B. McGillivray, Computational methods for semantic analysis of historical texts, Routledge, 2020.

[163] V. Iyer, M. Mohan, Y.R.B. Reddy and M. Bhatia, A Survey on Ontology Enrichment from Text (2019).

[164] M.N. Asim, M. Wasim, M.U.G. Khan, W. Mahmood and H.M. Abbasi, A survey of ontology learning techniques and applications, *Database* **2018** (2018). doi:10.1093/database/bay101.

[165] P. Buitelaar, P. Cimiano and B. Magnini, Ontology Learning from Text: An Overview, in: *Ontology Learning from Text: Methods, Evaluation and Applications*, Vol. 123, IOS Press, 2005, pp. 3–12.

[166] P. Cimiano and J. Volker, Text2Onto. A Framework for Ontology Learning and Data-driven Change Discovery, *Natural language processing and information systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15 - 17, 2005; proceedings. Lecture notes in computer science, 3513. Montoyo A, Munoz R, Metais E (Eds); Springer: 227-238.* (2005).

[167] D. Faure and C. Nédellec, Asium: Learning subcategorization frames and restrictions of selection (1998).

[168] X. Jiang and A.-H. Tan, CRCTOL: A semantic-based domain ontology learning system, *Journal of the American Society for Information Science and Technology* **61**(1) (2010), 150–168–. doi:10.1002/asi.21231.

[169] E. Drymonas, K. Zervanou and E.G.M. Petrakis, *Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System*, in: *Natural Language Processing and Information Systems*, C.J. Hopfe, Y. Rezgui, E. Métais, A. Preece and H. Li, eds, Lecture Notes in Computer Science, Vol. 6177, Springer Berlin Heidelberg, 2010, pp. 277–287–. ISBN 978-3-642-13880-5. doi:10.1007/978-3-642-13881-2_29. http://link.springer.com/10.1007/978-3-642-13881-2_29.

[170] R. Navigli and P. Velardi, Learning domain ontologies from document warehouses and dedicated web sites, *Computational Linguistics* **30**(2) (2004), 151–179.

[171] A. Oliveira, F.C. Pereira and A. Cardoso, Automatic Reading and Learning from Text, in: *Proceedings of the International Symposium on Artificial Intelligence (ISAI)*, 2001.

[172] U. Hahn and K. Schnattinger, Towards text knowledge engineering, *Hypothesis* **1**(2) (1998).

[173] H.-P. Zhang, Q. Liu, X.-Q. Cheng, H. Zhang and H.-K. Yu, Chinese lexical analysis using hierarchical hidden Markov model, *SIGHAN '03: Proceedings of the second SIGHAN workshop on Chinese language processing* **17** (2003), 63–70–.

[174] X. Zou, N. Sun, H. Zhang and J. Hu, *Diachronic Corpus Based Word Semantic Variation and Change Mining*, in: *Language Processing and Intelligent Information Systems*, M.A. Kłopotek, J. Koronacki, M. Marciniak, A. Mykowiecka and S.T. Wierzchoń, eds, Lecture Notes in Computer Science, Vol. 7912, Springer Berlin Heidelberg, 2013, pp. 145–150–. ISBN 978-3-642-38633-6. doi:10.1007/978-3-642-38634-3_16. http://link.springer.com/10.1007/978-3-642-38634-3_16.

[175] J.M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM* **46**(5) (1999), 604–632–.

[176] J.B. MacQueen, *Some methods for classification and analysis of multivariate observations*, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1967, pp. 281–297–. https://projecteuclid.org/euclid.bsmsp/1200512992.

[177] G.D. Rosin and K. Radinsky, Generating Timelines by Modeling Semantic Change, in: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Association for Computational Linguistics, 2019, pp. 186–195–. doi:10.18653/v1/K19-1018. https://www.aclweb.org/anthology/K19-1018.

[178] J.A. Gulla, G. Solskinnsbakk, P. Myrseth, V. Haderlein and O. Cerrato, SEMANTIC DRIFT IN ONTOLOGIES, in: *WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies*, Vol. 2, 2010.

[179] I. Augenstein, S. Padó and S. Rudolph, *LODifier: Generating Linked Data from Unstructured Text*, in: *The Semantic Web: Research and Applications*, E. Simperl, P. Cimiano, A. Polleres, O. Corcho and V. Presutti, eds, Lecture Notes in Computer Science, Vol. 7295, Springer Berlin Heidelberg, 2012, pp. 210–224–. ISBN 978-3-642-30283-1. doi:10.1007/978-3-642-30284-8_21. http://link.springer.com/10.1007/978-3-642-30284-8_21.

[180] K.T. Frantzi and S. Ananiadou, The C-value/NC-value domain-independent method for multi-word term extraction, *Journal of Natural Language Processing* **6**(3) (1999), 145–179. doi:10.5715/jnlp.6.3_145.

[181] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, in: *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, 1986, pp. 24–26–. https://dl.acm.org/doi/10.1145/318723.318728.

[182] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Linguistic Linked Data in Digital Humanities*, in: *Linguistic Linked Data. Representation, Generation and Applications*, 1st edn, Springer International Publishing, 2020. https://www.springer.com/gp/book/9783030302245.

[183] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, Linguistic linked open data cloud, in: *Linguistic Linked Data*, Springer, 2020, pp. 29–41.

[184] I. Maks, M. van Erp, P. Vossen, R. Hoekstra and N. van der Sijs, Integrating diachronous conceptual lexicons through linked open data, DHBenelux, 2016.

[185] J. Noordegraaf, M. van Erp, R. Zijdeman, M. Raat, T. van Oort, I. Zandhuis, T. Vermaut, H. Mol, N. van der Sijs, K. Doreleijers, V. Baptist, C. Vrielink, B. Assendelft, C. Rasterhoff and I. Kisjes, Semantic Deep Mapping in the Amsterdam Time Machine: Viewing Late 19th- and Early 20th-Century Theatre and Cinema Culture Through the Lens of Language Use and Socio-Economic Status, 2021, Accepted for publication.

[186] K. Depuydt and J. De Does, The diachronic semantic lexicon of dutch as linked open data, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Paris, France*, 2018.

[187] S. Tittel and F. Gillis-Webber, Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French, in: *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*, Lexical Computing, 2019, pp. 547–569.

[188] E. Vanhoutte, An Introduction to the TEI and the TEI Consortium, *Literary and linguistic computing* **19**(1) (2004), 9–16.

[189] A. Barbado, V. Fresno, Á.M. Riesco and S. Ros, DISCO PAL: Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels, *arXiv preprint arXiv:2007.04626* (2020).

[190] J.M. van der Zwaan, I. Maks, E. Kuijpers, I. Leemans, K. Steenbergh and H. Roodenburg, Historic Embodied Emotions Model (HEEM) dataset, Zenodo, 2016. doi:10.5281/zenodo.47751.

[191] I. Leemans, E. Maks, J. van der Zwaan, H. Kuijpers and K. Steenbergh, Mining Embodied Emotions: A Comparative Analysis of Bodily Emotion Expressions in Dutch Theatre Texts 1600-1800', *Digital Humanities Quarterly* **11**(4) (2017).

[192] A.F. Khan, Towards the Representation of Etymological Data on the Semantic Web **9**(12) (2018), 304, Publisher: MDPI AG. doi:10.3390/info9120304.

Appendix

Table 2. Main theoretical approaches surveyed in S 3

Knowledge-oriented	***Language-oriented***
Charting the history of political and social concepts [15]	Semasiological vs. onomasiological mechanisms of semantic change in lexical semantics [24]
Formal description of conceptual change implying a “core” and a “margin” [16]	Semasiological vs. onomasiological mechanisms of semantic change in cognitive linguistics and diachronic lexicology [25]
Defining the meaning of a concept in terms of “intension, extension and labelling” [11]	Stability and univocity principles vs. sociocognitive approaches to understand world and language change in terminology [28]
Model-based approach to the “history of ideas or concept drift” [20]	Diachronic change in the layer of pragmatics [29]
Describing semantic change, semantic drift, concept drift in relation to ontology change [17]	Discourse-historical approach (DHA) and the principle of "triangulation" [38]

Table 3. Main LL(O)D formalisms and resources surveyed in S 4 and S 7

Models	OntoLex-Lemon [41] Temporal RDF [55]; RDF-star
Approaches	Etymology modelling [48, 49, 192] Perdurantist modelling [57] OWL-based temporal reasoning [62]
Resources	LiLa etymological lexicon [52] OWL-Time ontology [60]; LODE ontology; PeriodO gazetteer of periods LL(O)D cloud [183] Linghub Diachronic semantic lexicon of Dutch [186]

Table 4. Main NLP methods for diachronic analysis surveyed in S 5

NER, NED, NEL	NER: rule-based [112–114]; unsupervised statistical [115]; machine learning [116–118]; deep learning [119–123] Time-aware NED, NER [124, 134] LL(O)D-based NEL [128–131]
Word embedding	Unsupervised [143]; hyperbolic [144, 145] LSTM-based [146]; detecting paradigmatic and syntagmatic shifts [147]
Transformer-based	BERT [149]; ELMo [150]; XLNet [151] Unsupervised [152]; clustering [153]
Topic modelling	SCAN [66]; topics over time LDA [156] Hierarchical Dirichlet [64, 65] LDA-based [159]

Table 5. Main NLP applications for generating (diachronic) ontological and linked data structures surveyed in S 6

Learning diachronic constructs	Ontologies [14, 144] Timelines [177] Concept signatures [178]
Learning ontologies and producing linked data	OntoGain [169] CRCTOL [168] TextToOnto [166]
Extracting information and linking entities	LODifier [179]
Converting to linked data formats	CoW, cattle [4] LLODifier [182]