

WESEARCH: LANGUAGE TECHNOLOGY FOR THE WEB

1 Relevance of the Proposed Research

Human Language Technology (HLT) is the interdisciplinary sub-field of computer science that aims to enable computers to ‘make sense’ of human language, for example to extract structured information from unstructured text, to determine the sentiment of a product review, to translate from one language into another, or to organize and retrieve Internet content on the basis of *abstract meaning*, rather than just keywords. As such, language technology is a key enabler to next-generation, user-centric ICT services and a link between technology R&D and the humanities.

The proposed project sets out to enable novel Web technologies in the broad realm of *social networking*. In user-centric information exchange and peer-to-peer interaction, so-called *user-generated content* (UGC) often is at the core of social networking and already accounts for a large proportion of Internet traffic. There are many facets of UGC, comprising for example audio and video sharing, community resources like Wikipedia, user forums and bulletin boards, professional or personal blogs, and of course iconic services like MySpace, facebook, and Twitter. Human language is the ‘fabric’ of the Web, and this project targets a representative sample of *textual* UGC. The focus here is on language technology to automatically assign formal representations of meaning to unstructured content (in English and Norwegian *Bokmål*¹)—a general purpose process dubbed *semantic parsing*. In the gradual transition from the Web 2.0 paradigm towards the Semantic Web (or Web 3.0), semantic parsing will be an essential tool.

The explosion in user-generated content has prompted an extensive interest in the field of text mining and in particular in automatic methods for tapping into the so-called ‘wisdom of the masses’: opinions and sentiments published on-line. In HLT, recent years have seen a growing interest in accessing the knowledge implicit in these massive information sources, so-called *sentiment analysis* or *opinion mining*. Recent research, however, shows that intelligent opinion mining requires a deeper analysis of the data (see § 2.1 below), in particular analyses which to a larger extent reflect the meaning of human language. A full semantic sentiment analysis application is beyond the scope of a single project, but it provides a technology vision characteristic of use scenarios made possible by Web-scale semantic parsing.

The project combines the horizontal theme of social networking with the vertical pillar of user interfaces, information management, and software technology. Quite generally, ‘intelligent’ access to steadily growing volumes of on-line content in the so-called knowledge society demand language-enabled ICT services. The Semantic Web is a powerful vision, but to move from predominantly unstructured Web content to rich semantic annotation, we will require the ability to automatically relate human language content to structured, formal semantics. In addressing a broad variety of user-generated content, the project facilitates integration across services and resources. More importantly, information retrieval today confines itself to matching keywords, emphasizing *recall* over *precision*, with poor results against dynamic, low-density content (user forums, for example). Next-generation services for intelligent information access need to overcome these limitations and scale better to structured queries or advanced retrieval problems (e.g. searching specifically for *critical* reviews).²

Finally, the project also addresses the vertical pillar on social, economic, and cultural challenges and opportunities, if mostly in terms of expected mid- and long-term impact. Multilingualism—presence and use of different languages side-by-side—is a challenge of growing importance: less than one third of Web content today is English (compared to about half around the year 2000). The EU embrace multilingualism as an economic and cultural strength, and future Internet technology will demand services that transparently work across language barriers.

¹However, main project results will be language-independent methodological and technological advances, which will be directly applicable to other languages.

²Start-up companies like Hakia (<http://www.hakia.com>), NetBase (<http://www.netbase.com>), or PowerSet (<http://www.powerset.com>; now a division of Microsoft) are working to commercially establish ‘semantic search’, each with its own proprietary technology. These efforts demonstrate, on the one hand, that the proposed project is a timely and economically relevant R&D initiative. On the other hand, foundational research based on generally available methods and technology is required for long-term scientific progress in this area.

2.1 Background and Status of Knowledge

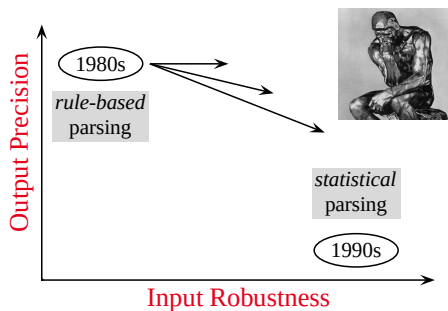


Figure 1: Dimensions of comparison.

Parsing human language is the process of analyzing grammatical structure, i.e. working out *who did what to whom?* Arguably parsing is a mandatory step towards interpretation and has always been among the most central techniques in the relatively short history of human language technology. Mirroring developments in other fields, notably Artificial Intelligence, approaches to parsing human language originally were anchored in formal grammars, hand-built systems of linguistic rules. Due to the cost of development and computational complexity involved, rule-based parsers typically do not scale easily to free text. Increased adoption of statistical methods and the availability of large corpora of text annotated with grammatical structure since the mid-1990s led to a surge of parsers trained automatically on annotated corpora, notably the Penn Treebank (PTB; Marcus, Santorini, & Marcinkiewicz, 1993). Such statistical parsers can achieve good results on texts similar to the training material, but often they are difficult to adapt to different types of input (Gildea, 2001). Figure 1 gives a schematic view of evolution along the two dimensions, suggesting a classic trade-off: the ability to successfully process unrestricted inputs comes at the expense of *output precision*, by which we refer to both the granularity (or amount) of information available and the average correctness of outputs.

Emphasizing the vertical dimension of comparison, parsing approaches are often classified as ‘deep’ vs. ‘shallow’. These terms, however ubiquitous in the scientific discourse, remain at best poorly defined. A prototypical ‘shallow’ parser will not explicate the difference in use of the preposition *in* in examples like *Cisco is interested in Tandberg.* vs. *Tandberg is headquartered in Oslo.* Conversely, a ‘deep’ parser would be expected to recover the underlying relations in a sentence like *Cisco, CNN reports, sets out to acquire Tandberg in December.* Here, the network maker is the agent of both the aiming (the semantics of the multi-word *set out*) and the acquiring—see Figure 3 below. The arrows in Figure 1 symbolize recent and ongoing developments, with contemporary R&D now seeking to combine the strengths of both worlds. Hybrid paradigms combine in-depth linguistic analysis with sophisticated statistical methods. A third dimension of variation is the cost of computation. More in-depth analysis and higher granularity of output information makes the parsing problem harder, typically enlarging the search space of candidate solutions. With per-sentence parsing times varying between several milliseconds and a few seconds, algorithmic efficiency remains a technological barrier to the large-scale deployment of deeper analysis.

There is a wide range of user-centric HLT applications to Web content that motivate R&D towards semantic parsing as a general purpose technology. Areas such as entity and relation extraction, ontology learning, social network analysis, and others already experience limitations imposed by a lack of reliable, large-scale analysis of grammatical structure. Returning to our application vision of opinion mining over user-generated content, parsing has been demonstrated to significantly improve sentiment analysis (Choi, Breck, & Cardie, 2006). A sentiment is a relation in which a *source* has an evaluative *judgment* regarding a *target*. Sentiment analysis traditionally is applied at the document level, e.g. determining whether a product review is broadly positive or negative.

Recently, however, the recognition of sentiment at the level of individual utterances (typically a one-sentence statement) has received more attention. This involves subtasks such as the recognition of opinion holders (entities), determining the strength and polarity of opinion expressions, and the joint extraction of opinion entities and their relations (Choi et al., 2006). It has been shown that the use of relational linguistic information is important in particular in expression-level sentiment analysis (Bunescu & Mooney, 2005; Choi et al., 2006; inter alios), and more generally that structural choices directly influence the perceived sentiment of an utterance (Greene & Resnik, 2009). Standard ‘bag-of-words’ approaches in information retrieval, however, are agnostic to structural relations in human language and essentially use identical representations for the sentences *Cisco acquired Tandberg* vs. *Tandberg acquired Cisco*. Consider two examples from technology blogs:

- (1) Google isn't universally loved by newspaper execs and other content creators.³
- (2) Despite the overall attractiveness of the iPhone, it lacks too many vital features to replace the BlackBerry as the corporate weapon of choice.⁴

Example (1) above invokes the so-called passive construction (which is frequent in many languages), where the agent phrase *newspaper execs and other content creators* is not expressed as the syntactic subject (which in English tends to occur sentence-initially), but rather as an argument of the preposition *by*. The semantic 'object' of the *love* relation (i.e. Google), on the other hand, occupies the position of the syntactic subject. Furthermore, the negation expressed by the contracted auxiliary *isn't* crucially takes scope over the complete sentiment expressed by the main verb and its arguments. Conversely, example (2) above invokes a comparison of two smartphones and relies on a contrasting relation (provided by *despite*) between two independent statements. Without working out the grammatical structure of the utterance as a whole, including resolution of the pronoun *it* as referring to the iPhone and recognition of inherent negation in words like *lack*, it would be impossible to say which device is favored. Such structural relations are captured by semantic analysis (see Figure 3 below for an example representation).

There is insufficient knowledge for empirical comparison of competing approaches. Recent parsing R&D for English has near-exclusively focused on a subset of the syntactic information available in the PTB and its narrow subject matter: Wall Street Journal articles from the late 1980s. Extant parser evaluation metrics (of which some have been used widely) thus fail to provide insights into what it means for linguistic analysis to be 'deep' to some degree. Distinct approaches offer seemingly comparable performance on this limited task (Miyao, Sagae, & Tsujii, 2007). In contrast, user-generated Web content presents very different challenges, stemming in part from informal and more creative language use, and in part from 'noise' inserted in electronic communication. Likewise, no studies exist that systematically correlate linguistic adequacy with non-linguistic properties—robustness to diverse inputs, processing efficiency, and adaptability across domains, for example.

Scalable (semantic) parsing technology is too expensive to develop for a single player, corporate or academic. The interdisciplinary nature of the task, combining linguistics and computer science (and, in principle at least, also cognitive psychology), calls for the incremental and sustained development of formal theory, computational grammars and annotated corpora, as well as carefully implemented technology. Therefore, a long-term perspective, collaborative development, focus on task-, genre-, and domain-adaptable approaches, and the exchange and reuse of knowledge and resources are widely viewed as prerequisites to broader use of parsing in next-generation ICT solutions.

2.2 Problem Statement, Hypotheses, and Choice of Methods

The proposed research takes as its point of departure two fundamental assumptions: (a) semantic parsing of large and diverse volumes of human language content will be of growing importance to the evolution of the Internet of the Future; and (b) scientific progress to date and long-term uptake are hindered by a vague problem definition, lack of general purpose technology, the unsolved problem of domain (and genre) adaptation, and insufficient interface standardization and documentation. These need to be approached in close cooperation with key international players.

At the same time, we believe that linguistic theory and computational parsing have matured to a point where large-scale application to Web content is within reach. However, there are substantive scientific and technological challenges that need to be addressed to reach this goal. These pertain to the (search for equilibrium along the) three dimensions of variation discussed earlier: robustness, precision, and efficiency. Besides pushing ahead in all three directions, it is important to determine (and document) precisely what state-of-the-art technology can and cannot deliver—for different types of inputs and parser configurations—and what the trade-offs are along the three dimensions. Furthermore, to establish semantic parsing as a general purpose technology, there are methodological shortcomings to overcome. These include techniques for adaptation and reuse across different domains, genres, and tasks, as well as for flexible parametrization and configuration, choosing an optimal balance of robustness, precision, and processing cost for a specific use scenario.

³Taken from The Copywriter Underground, accessible through <http://tinyurl.com/wesearch01>.

⁴Taken from TechCrunch, accessible through <http://tinyurl.com/wesearch02>.

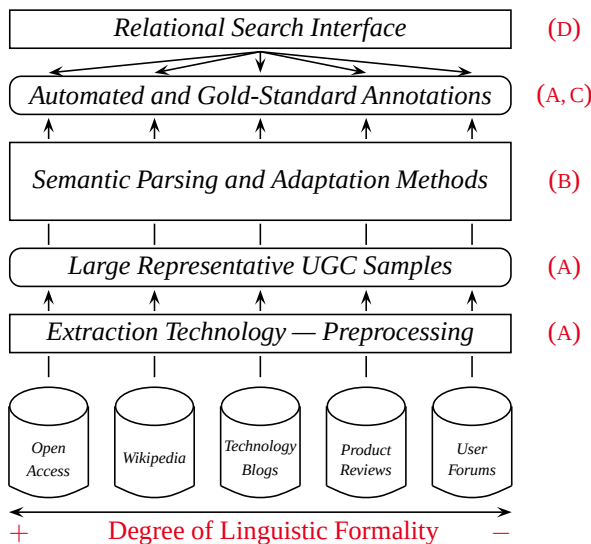


Figure 2: Materials, resources, and technology.

ative success in a multi-dimensional grid of linguistic and computational metrics; and (c) anchor its results in a collaborative, multi-national effort to better define the semantic interface to grammatical analysis. If successful, the proposed research will advance the state of the art in language technology in several important areas, including a sharper definition for the task of semantic parsing, greatly improved parsing and adaptation techniques, detailed knowledge about the effects of relevant choices, and a community process towards a de-facto interface standard. As an important by-product, the project will compile novel UGC corpora for English and Norwegian.

The proposed research is organized in four intertwined tracks, with one core effort—track (B): semantic parsing—and three smaller support activities. Figure 2 gives a schematic overview, and the following paragraphs summarize objectives, methods applied, and expected results for each track.

(A) Resource Creation Language technology R&D is inherently data-driven: most development and experimentation are organized around electronic language data. The project will investigate (and advance) the fitness of semantic parsing technology for diverse types of user-generated Web content. But very few existing corpora to date exemplify the linguistic complexity and variation of this kind of language. Hence, this track will develop and apply tools to harvest and preprocess substantial samples of on-line content, in both languages. Materials will be drawn from: hard- and software user forums, reviews of consumer electronics, science and technology blogs, Wikipedia, and open-access research literature.⁵ This selection reflects linguistic variation (ranging from edited, formal language to dynamic and informal language), establishes the broad domain of information technology, and connects to prior and ongoing work in the field. *Prima facie* maybe unexpectedly, there are unsolved scientific questions in this task related to (a) the interface of text-level markup and grammar, as well as to (b) ‘noise’ characteristic of some channels of on-line communication.

Markup elements like headings, bulleted lists, or italics can indicate specialized grammar, foreign words, or a so-called use – mention contrast (as in ‘the preposition *by*’ in § 2.1 above). Linguistically relevant markup should be available to parsing, while purely design-oriented markup elements must be suppressed. Different markup standards are in use (e.g. (X)HTML, L^AT_EX, Wiki Syntax, and various ASCII conventions), but the grammatical rules of English (or another language) should rather refer to relevant text-level properties in logical terms—for example an abstract *italics* property. The project will extend earlier work by Ytrestøl, Flickinger, & Oepen (2009) and identify relevant markup elements in various types of Web content and define a suitably *general markup language*, used to annotate parser inputs. In joint work with the University of Melbourne, we will extend the open-source SiteScraper framework (Penman, 2009) for extraction of linguistic content and relevant meta information from the above channels. The resulting text collections will be automatically pre-processed (including sentence boundary detection) to provide inputs for tracks (B) and (C); where

⁵For this latter, most formal type of UGC, we draw on earlier work at Oslo University (UiO) and DFKI Saarbrücken, targeting the Norwegian Open Research Archives (NORA) and ACL Reference Corpus (ARC), respectively.

In summary, the proposed research is technology-driven and empirical in nature. Our main hypothesis here is that an insightful synthesis of existing knowledge and current best practices—across approaches, frameworks, and languages—will afford sufficient incremental advances in methodology and technology to enable Web-scale semantic parsing. A subsidiary hypothesis relates to the interaction of theory and technology development. We believe that adaptability, scalability, and long-term scientific progress in parsing technology can only be obtained through the combination of leading linguistic and computer science expertise—reflecting the interdisciplinary nature of the problem. The project seeks to (a) validate this proposition through the development of semantic parsing technology for a large and representative selection of user-generated Internet content; (b) quantify relative

possible, materials will be chosen that allow the redistribution of these corpora to the general public.

When extracting text from PDF files, scraping a Web site, or analyzing the archives of a user forum, there is a certain level of non-linguistic ‘noise’, for example hyphenation at line breaks, intervening navigation elements or advertisement, and other artifacts of recovering linguistic content by ‘reversing’ display presentation. Working together with DFKI, the project will investigate automated techniques for the detection, removal, or correction of such artifacts. Processing noisy language data is an HLT task of emerging importance,⁶ and the data and methods developed in this track will therefore create immediate international interest.

Finally, for project-internal evaluation and interface corroboration, parts of our UGC corpora will be annotated with gold-standard syntactic and semantic structures, resulting in what is commonly referred to as *treebanks* (even though semantic structures need not take the form of trees). Although such annotation is generally expensive to construct, the project setup facilitates the use of semi-automatic treebanking techniques. The Redwoods approach (for English: Oepen, Flickinger, Toutanova, & Manning, 2004) and its TrePil adaptation to Norwegian (Rosén, Smedt, Dyvik, & Meurer, 2005) allow the construction of comparatively detailed treebanks at greatly reduced cost, leveraging the parsers from track (B) and providing a specialized annotation tool to identify the correct analyses among the space of possible parser outputs. Following an initial round of parser adaptation to the various texts, treebanks of sizable subsets of the data will be prepared in the second year of the project, and then be maintained through semi-automated updates until project completion.

(B) General Purpose Semantic Parsing The project builds on decades of prior work on *computational grammars* by project partners: the English Resource Grammar (ERG; Flickinger, 2000) continuously developed at Stanford University since 1993, and NorGram (Dyvik, 1999) under development at Bergen University (UiB) since 1999. These grammars implement two of the arguably most influential grammatical theories for HLT and run on specialized software: the PET (Callmeier, 2002) and XLE (Maxwell and Kaplan, 1995) parsing systems, respectively. These resources embody the ‘deepest’ broad-coverage approaches to parsing available today and have long been key elements of multi-national collaboration, the DELPH-IN⁷ and ParGram⁸ networks, respectively. Building on successful prior cooperation between UiO, UiB, and Stanford (in the NFR-funded LOGON and HandOn projects), our setup presents a rare opportunity for cross-framework and cross-language investigation, which will greatly add to the generality and credibility of project results.

$\langle h_1,$
 $\left\{ \begin{array}{l} h_7:\text{named}(x_5, \text{CNN}), h_8:\text{report.v}(e_2, x_5, h_{10}), \\ h_{15}:\text{named}(x_{13}, \text{Cisco}), h_{16}:\text{set+out.v.aim}(e_{17}, x_{13}, h_{18}), \\ h_{19}:\text{acquire.v}(e_{20}, x_{13}, x_{21}), h_{26}:\text{named}(x_{21}, \text{Tandberg}), \\ h_{19}:\text{in.p.temp}(e_{20}, x_{23}), h_{26}:\text{mofy}(x_{23}, 12) \end{array} \right\}$
 $\{ h_1 =_q h_8, h_{18} =_q h_{19}, h_{10} =_q h_{16} \} \rangle$

Figure 3: Sample (simplified) semantic representation.

The ERG and NorGram output meaning representations in the widely used Minimal Recursion Semantics framework (MRS; Copestake, Flickinger, Pollard, & Sag, 2005). Figure 3 shows a simplified MRS for our earlier example: *Cisco, CNN reports, sets out to acquire Tandberg in December*. Ignoring formal detail, structural relations hold between logical variables,

e.g. the entity x_5 partaking in both the naming (as *CNN*) and reporting relations, or the acquiring event e_{20} being temporally restricted to the 12th month of the year (‘mofy’). As required in § 2.1 above, the shared agent of the aiming and acquiring events is captured by variable identity (x_{13}), and the equality of ‘handle’ variables h_{10} and h_{16} embeds the ‘setting out’ as the semantic object of the reporting. Such will be the general form of interface representations for the semantic parsers.

While the ERG has been applied to informal communication before (spoken dialogues and email), NorGram development so far has focused on edited language. The project requires a limited amount of grammar extension and adaptation, particularly adding analyses of new UGC-specific grammatical constructions and non-sentential utterances. The core of the proposed research, however, will focus on technology improvements and experimentation geared at increased input robust-

⁶For example, noisy data is the thematic focus of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), for which the proposer serves as one of 18 Area Chairs.

⁷Deep Linguistic Processing with HPSG Initiative; <http://www.delph-in.net>.

⁸Parallel Grammars; <http://pargram.b.uib.no>.

ness, output precision, parsing efficiency, and techniques for adaptation and configuration.

First, the ERG parser now standardly applies a so-called Part of Speech (PoS) tagger for pre-processing its input; PoS tagging is a (statistical) sequence classification process assigning coarse-grained lexical categories to input tokens—e.g. noun, verb, or adjective—prior to full grammatical analysis. For words unknown to the ERG lexicon, PoS tags are used to robustly provide under-specified, fall-back lexical entries.⁹ However, seeing massive lexical ambiguity in both English and Norwegian (forms like *dances* or *danse* can be nominal or verbal), reliable PoS tags could in principle be used to prune the search space of the semantic parser, or to guide the statistical ranking of alternate analyses. Thus, besides increased robustness, novel methods for tighter integration of PoS tagging and parsing will likely yield improvements in output precision and parsing efficiency.

An extension of PoS tagging is so-called supertagging, sequence classification of parser inputs with much more detailed lexical categories. Where standard PoS tags make a few dozen distinctions, supertags tend to number in the hundreds or thousands. Albeit abstractly similar to using PoS tags, supertagging is a more difficult preprocessing task, and making optimal use of supertags in semantic parsing an unsolved problem. A recent dissertation at DFKI, on parsing with the ERG, (Dridan, 2009; co-supervised by the proposer) suggests potential for substantial improvements in robustness, precision, and efficiency. Preliminary experiments, however, remain isolated along one dimension of variation at a time and limited to formal English. The project will address remaining issues in this work, develop new techniques for combining supertagging with the ERG and NorGram, and seek to validate these in multi-dimensional experiments.

Another area of parser hybridization for increased robustness to out-of-scope inputs is so-called *robust parsing*. Here, the project will experimentally contrast the approach dominant in parsing with the XLE (heuristic integration of parse fragments) with the work of Zhang & Kordoni (2008) (a PCFG fall-back and robust semantic construction), seek a synthesis of strong points of both techniques, and document robustness vs. precision trade-offs when parsing user-generated content. Resulting algorithmic enhancements will need to be compatible with the approach to improved XLE parsing efficiency of Cahill, Maxwell, Meurer, Rohrer, & Rosén (2008).

Second, disambiguation is the task of selecting the ‘correct’ (most probable) parse(s) from the space of possible analyses delineated by the grammars. Whether rule-based at its core or not, statistical disambiguation is a central component of all computational parsing nowadays—typically training high-dimensional Maximum Entropy or SVM machine learning models from treebanks (Abney, 1997; Johnson, Geman, Canon, Chi, & Riezler, 1999; Toutanova, Manning, Flickinger, & Oepen, 2005; inter alios). For inputs of about 17 words average length, current ERG parse disambiguation ranks the correct analysis highest in less than sixty per cent of cases, even though it is among the top-ten candidates according to the model in nine of ten cases. Fujita, Bond, Oepen, & Tanaka (2007) demonstrate the efficacy of semantic properties in parse disambiguation (for Japanese), but current models used with the ERG and NorGram do not exploit such information. For the simpler case of context-free syntactic parsing, Charniak & Johnson (2005) suggest a range of relevant properties, applied in a two-stage disambiguation setup. The proposed research will synthesize these approaches into a uniform framework, seeking to substantially improve parse selection accuracy for the ERG and NorGram through semantic and additional non-local properties, and to combine the benefits of improved precision with (super)tagging and early pruning in a cascaded approach to disambiguation.

Finally, parse disambiguation models are standardly trained in a *supervised* manner—from annotated training data—which inhibits adaptability across domains and genres. Plank & van Noord (2008) and Rimell & Clark (2008) suggest semi-supervised techniques, assuming only small quantities of domain-specific annotation. The project will adapt these and similar techniques to semantic parsing with the ERG and NorGram, experimentally contrast these approaches on various types of UGC, and seek a further reduction in adaptation cost through the combination with so-called *self training* (McClosky, Charniak, & Johnson, 2008), a technique which Dridan (2009) suggests can be especially beneficial in supertagging.

⁹The ERG and NorGram include comprehensive grammatical lexicons (of approximately 40,000 and 80,000 entries, respectively), but unknown word handling is nevertheless essential for robustness to running text.

(C) Interface Corroboration Formal semantic representations as a generic interface to parsing (abstractly, a linguistic API) are subject to constraints from multiple disciplines: (computational) linguistics capitalizes on representational (or explanatory) adequacy and compositional compatibility with syntax and morphology; formal logic has its focus on mathematical properties and support for logical inference; application development, finally, emphasizes more practical requirements—sufficiently detailed, yet easy-to-comprehend representations that are stable over time. Reconciling such different points of view remains a major scientific challenge and a traditional barrier to wider uptake of semantic parsing for HLT applications. The project cannot fully resolve these issues, but it will seek to become a catalyst (and ‘clearing house’) for a recent surge of international interest in harmonizing and documenting semantic representations, for improved application support.

The different types of stakeholders are all represented among project partners, and additional key players (from DELPH-IN, ParGram, and beyond) will be invited to participate in a longer-term community effort towards semantic standardization. This effort will take a data-centered approach, based on a careful manual selection (of at most a few hundred utterances in English and Norwegian) from UGC samples compiled in track (A), combined with a ‘catalog’ of frequent semantic phenomena. The data will be released publicly, challenging interested research groups to submit structural semantic annotation (manually constructed analyses or hand-corrected parser outputs). In a first round of investigation, these representations will be aligned (at the level of sub-string positions) and provide the basis for an in-depth, by-invitation workshop held by the project in its second year.¹⁰

Reflecting on this input, the project will prepare gold-standard semantic analyses for this data (inevitably focusing on English, but maintaining a cross-linguistic perspective), accompanied by descriptive summaries of core analyses and the underlying motivations—which, in some cases, may just point out stand-in solutions where a fully satisfactory analysis is not yet available. Annotation and documentation will be made available to external collaborators in a Wiki-based interface (linking to an on-line demonstrator for the parsers of track (B), for user experimentation); a phase of community critique and discussion will prepare a second topical workshop organized by the project. The final revision of gold-standard analyses and documentation and a more general summary of formal and application-oriented desiderata for computational semantics will comprise the final results of this track, serving as greatly improved interface documentation and as a foundation for mid- and long-term standardization.

(D) Application Showcase Our application vision of sentiment analysis from semantic parses provides a longer-term goal. For this project, we propose an application that can showcase the utility of Web-scale structural semantics directly at the level of MRS representations: content retrieval taking advantage of semantic relations. Imagine a user searching for smartphones that support the Scalable Vector Graphics (SVG) Web extension. In part owed to ambiguity of the terms *support* and *SVG*, a keyword-based Internet search yields unsatisfactory results. There are relevant entries from Wikipedia, technology blogs, and review sites, of course—but few make it into the top few pages of standard search engine results. In contrast, posing a structured query against semantically parsed Web content would retrieve fewer but more relevant matches, i.e. higher search precision at (potentially) reduced recall. In MRS, a relational query could essentially take the following form: $\{ \text{_smartphone.n}(x_1), \text{_named}(x_2, \text{_SVG}), \text{_support.v}(_, x_1, x_2) \}$.

Schäfer, Uszkoreit, Federmann, Marek, & Zhang (2008) present a search infrastructure based on MRS relations developed at DFKI, and this is the technology we plan to adapt for this track. DFKI is currently engaged in a closely related project (also using the ERG, PET, and MRS) and have agreed to the exchange of experience and software, and to jointly developing a comprehensive relational search showcase for ICT-related information. Reflecting the scale of tracks (A) and (B) above, the demonstrator will index many millions of IT-related documents (though not billions, as general Web search engines do), with a focus on English and enabling information access across UGC types and services. To make the showcase more attractive, recall can be increased through integration of query expansion (*cell phone* $\subset$ *smartphone* or *support* $\subset$ *implement*, for example) and coreference resolution of pronouns, both R&D activities already underway at UiO and DFKI.

¹⁰This work builds on an earlier collaborative effort, orchestrated in 2007 and 2008 by Cambridge University with UK funding, as well as extending an initiative first launched as a workshop on parser evaluation, co-organized by the proposer as part of the 2008 International Conference on Computational Linguistics, which attracted about 60 participants.

2.3 Project Plan, Management, Organization, and Collaboration

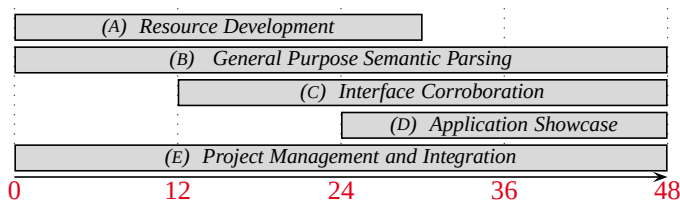


Figure 4: Sequencing and overlap of R&D tracks.

Tracks (A) to (D) will be executed in overlapping succession, as shown in Figure 4, with a lightweight management track (E).¹¹ Dependencies and sequencing requirements among the four tracks of investigation are spelled out in the grant application form, through main work packages and milestones, i.e. intermediate results that (a) enable subsequent work and (b) allow dissemination through scientific publication.

The first such milestone is a detailed work plan and schedule, due within three months after project kick-off, based on a planning meeting of all project partners receiving funding. Given the collaborative character of the proposed research and some as yet unknown external dependencies, it is difficult to finally determine the course of work beforehand.

Cohesion across tracks and collaboration among partners will be managed through repeated (internal) releases of a shared repository of project materials, i.e. the full collection of corpora, linguistic resources, and software involved. There will typically be three such snapshots per project year, and each will be prepared and implemented on the basis of a detailed release schedule.¹² The partners have worked together successfully before and build on existing routines and collaboration infrastructure, specifically a remote cooperation platform. The project as a whole will be coordinated at UiO by the proposer, with Helge Dyvik as a co-proposer managing R&D at UiB. The proposers contribute to all research tracks and project management and jointly take responsibility for the success of the project. Dyvik and Oepen represent partly complementary expertise: (computational) linguistics and language technology, respectively. UiO will be involved in all four tracks, focusing on English, and UiB in tracks (A) – (C), with a focus on Norwegian. Both sites will employ one (somewhat experienced) post-doctoral fellow over the full project duration, to participate in all scientific aspects of the project. One doctoral fellow funded by the project will work at UiO on one specific sub-problem of track (B), either robust parsing or semi-supervised domain adaptation. The thesis project will be defined as a part of the full semantic parsing pipeline, though not on a critical path.

The original ERG developers at Stanford will be sub-contracted for the equivalent of one person year, spread out over the first three years of the proposed research. Stanford primarily provides in-depth linguistic expertise and serves in an essential supporting role for grammar adaptation, error analysis, and treebanking. Furthermore, Stanford will contribute centrally to track (C), specifically to the selection and annotation of examples, the phenomena catalog, and the documentation of analyses. These activities tie into ongoing parallel work at Stanford (and prior collaboration with Cambridge) and thus all parties involved expect strong effects of synergy and mutual benefits. Cambridge University, DFKI, and Melbourne University participate in the project using their own funding. All three are actively working on large-scale semantic parsing and Web-scale applications. Cambridge is the main inventor of the MRS framework (jointly with Stanford), DFKI the original developer of the PET parser and relational search infrastructure, and Melbourne the creator of the Web site and user forum analysis technology to be used in the project. The sites have a prior track record of pairwise cooperation. Unfunded partners will contribute to the proposed research through mutually shared resources and technology, exchange of experience, synchronization of experiments and publications, and mutual visits and staff exchange. All partners will be invited to participate in bi-annual project meetings and the project-organized topical workshops.

Through participation in DELPH-IN and ParGram, the UiO and UiB project teams are in strong positions already to liaison effectively to the larger HLT community. The proposed research will help establish cutting-edge semantic parsing expertise at leading Norwegian research groups in our field—which will greatly increase international visibility and networking—as well as prepare the ground for technology transfer and industrial innovation—the foundations for novel, user-centric and language-enabled Web services and next-generation social networking.

¹¹The grant application form provides the exact schedule of deliverables and milestones

¹²Joint release planning and a review of routines will be an essential part of bi-annual multi-day meetings of researchers funded by the project, the proposers, and close affiliates where appropriate.

3.1 Compliance with Strategic Objectives

Since 2007, language technology is recognized as a focal research area by the UiO Faculty of Mathematics and Natural Sciences and the Department of Informatics (IFI).¹³ The proposer was appointed into a newly-created faculty position and is heading the HLT R&D area at IFI. A third faculty position in HLT is currently under review, and the faculty and department provide the group with four doctoral and one post-doctoral fellowships. With funding from NORA and NoTur, the group works closely with high-performance computing at UiO and utilizes the 5000-core TITAN cluster. The proposed project will help bundle ongoing activities at IFI and provide focus and structure: two of the doctoral fellows have closely related projects—near-deterministic parsing using the ERG, and MRS-based coreference resolution. The post-doctoral researcher works on statistical disambiguation. All three will be associated to the proposed project and contribute half of their time. Therefore, this initiative bears the potential of creating substantial results at comparatively moderate cost.

3.2 Relevance to Society

Beyond its contributions to the scientific understanding of human language and cognition, language technology brings the potential to transform information access and on-line interaction, as an enabler for next-generation, user-centric Web services and ‘intelligent’ IT systems more generally. The ability to semantically annotate the knowledge embodied in massive repositories of user-generated content will contribute to the development of new social networking paradigms and, ultimately, help advance an equal-opportunity knowledge society and technological innovation.

3.3 Environmental Perspectives

The proposed research has no immediate ramifications for the external environment.

3.4 Ethical Aspects

There are no (known) ethical issues relating to the proposed research or exploitation of results.

3.5 Gender Equality and Gender Perspectives

Language technology lags behind in gender balance. For all fellowships, appropriate efforts will be made to recruit woman experts (applying *svak kjønnskvotering*), with promising candidates available among recent graduates. The research team will put strong emphasis on creating a balanced environment where students and staff of both genders can collaborate comfortably and successfully.

4.1 Communication with Users and Dissemination of Results

As a research initiative, the project has as its primary ‘user base’ the scientific partners and larger research community. Dissemination of results, in this phase, will target scientific peers and the general public. The proposed research includes strong implementation and resource building components, and the vast majority of project results will be distributed, as they become available, with open-source licensing. This model is best-suited to the HLT world, as it assures ready dissemination, uptake, and longevity of resources. The repository of resources provided by the project will be a primary channel of outside communication.

In terms of scientific dissemination, conference presentations and publications in high-impact channels comprise an integral component of all work packages. The project team will aim to jointly publish a ‘position statement’, based on the exchange with scientific partners and results of the early work packages, and aiming to create awareness and visibility for the specific approach of the project. Interaction with international project partners, participants in project-organized workshop, and early adopters of project results will further enhance international visibility.

Finally, in terms of communication to the general public (including future industrial partners, application developers, and the taxpayer), the project Web site will provide a central repository of up-to-date information. The site will make available information at two levels, one general and jargon-free, another more specialized and technical.

¹³See <http://www.ifi.uio.no/research/groups/lms/lt.html> for additional background.

References

- Abney, S. P. (1997). Stochastic attribute-value grammars. *Computational Linguistics*, 23, 597–618.
- Bunescu, R. C., & Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 724–731). Vancouver, Canada.
- Cahill, A., Maxwell, J. T., Meurer, P., Rohrer, C., & Rosén, V. (2008). Speeding up LFG parsing using c-structure pruning. In *Proceedings of Grammar Engineering across Frameworks* (pp. 33–40). Manchester, UK.
- Callmeier, U. (2002). Preprocessing and encoding techniques in PET. In S. Oepen, D. Flickinger, J. Tsujii, & H. Uszkoreit (Eds.), *Collaborative language engineering. A case study in efficient grammar-based processing* (pp. 127–143). Stanford, CA: CSLI Publications.
- Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43th Meeting of the Association for Computational Linguistics* (pp. 173–180). Ann Arbor, MI.
- Choi, Y., Breck, E., & Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 431–439). Sydney, Australia.
- Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (2005). Minimal Recursion Semantics. An introduction. *Journal of Research on Language and Computation*, 3(4), 281–332.
- Dridan, R. (2009). *Using lexical statistics to improve HPSG parsing*. Doctoral dissertation, Saarland University, Saarbrücken, Germany.
- Dyvik, H. (1999). The universality of f-structure. Discovery or stipulation? The case of modals. In *Proceedings of the 4th International Lexical Functional Grammar Conference*. Manchester, UK.
- Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1), 15–28.
- Fujita, S., Bond, F., Oepen, S., & Tanaka, T. (2007). Exploiting semantic information for HPSG parse selection. In *Proceedings of the ACL Workshop on Deep Linguistic Processing* (pp. 25–32). Prague, Czech Republic.
- Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing* (pp. 167–202). Pittsburgh, PA.
- Greene, S., & Resnik, P. (2009). More than words: Syntactic packaging and implicit sentiment. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* (pp. 503–511). Boulder, CO.
- Johnson, M., Geman, S., Canon, S., Chi, Z., & Riezler, S. (1999). Estimators for stochastic ‘unification-based’ grammars. In *Proceedings of the 37th Meeting of the Association for Computational Linguistics* (pp. 535–541). College Park, MD.
- Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English. The Penn Treebank. *Computational Linguistics*, 19, 313–330.
- Maxwell III, J. T., & Kaplan, R. M. (1995). A method for disjunctive constraint satisfaction. In M. Dalrymple, R. M. Kaplan, J. T. Maxwell III, & A. Zaenen (Eds.), *Formal issues in Lexical-Functional Grammar* (pp. 381–401). Stanford, CA: CSLI Publications.
- McClosky, D., Charniak, E., & Johnson, M. (2008). When is self-training effective for parsing? In *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 561–568). Manchester, UK.
- Miyao, Y., Sagae, K., & Tsujii, J. (2007). Towards framework-independent evaluation of deep linguistic parsers. In T. H. King & E. Bender (Eds.), *Proceedings of Grammar Engineering across Frameworks* (pp. 238–258). Stanford, CA: CSLI Publications.
- Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. *Journal of Research on Language and Computation*, 2(4), 575–596.
- Penman, R. B. (2009). Web scraping made simple with SiteScraper. In *Proceedings of the 18th International World Wide Web Conference*. Madrid, Spain.
- Plank, B., & van Noord, G. (2008). Exploring an auxiliary distribution-based approach to domain adaptation of a syntactic disambiguation model. In *Proceedings of the COLING Workshop on Cross Framework and Cross Domain Parser Evaluation* (pp. 9–16). Manchester, UK.
- Rimell, L., & Clark, S. (2008). Adapting a lexicalized-grammar parser to contrasting domains. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 475–484). Honolulu, HI.
- Rosén, V., Smedt, K. D., Dyvik, H., & Meurer, P. (2005). TREPIL. Developing methods and tools for multilevel treebank construction. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories* (pp. 161–172). Barcelona, Spain.
- Schäfer, U., Uszkoreit, H., Federmann, C., Marek, T., & Zhang, Y. (2008). Extracting and querying relations in scientific papers on language technology. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 127–134). Marrakech, Morocco.
- Toutanova, K., Manning, C. D., Flickinger, D., & Oepen, S. (2005). Stochastic HPSG parse selection using the Redwoods corpus. *Journal of Research on Language and Computation*, 3(1), 83–105.
- Ytrestøl, G., Flickinger, D., & Oepen, S. (2009). Extracting and Annotating Wikipedia Sub-Domains. Towards a New eScience Community Resource. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories* (pp. 185–197). Groningen, The Netherlands.
- Zhang, Y., & Kordoni, V. (2008). Robust parsing with a large HPSG grammar. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 1888–1893). Marrakech, Morocco.