1 Abstract

The goal of TREiL is to show that a few simple 1technological tools, borrowed from web design, computational linguistics and the tradition of hands-on models used for the popularization of 'hard sciences', can greatly improve the end-to-end process of doing theoretical linguistics. Building on previous European projects, the project aims to carry out the following tasks:

(T1) develop better methods for linguistic data collection via the web, with focus on oral data and minority language, and a community-building approach which can increase the participant's motivation;

(T2) explore a novel methodology to compare grammaticality judgment from human speakers, partly obtained from T2, with the latest corpus-based computational models of language (based on recurrent neural networks). Since we can be sure that such models have no pre-wired, language-specific learning bias, they can be used to test one of the fundamental questions in theoretical linguistics: to what extent there is an innate, language-faculty-specific "universal grammar".

(T3) increase the outreach of theoretical linguistics, which unlike other scientific disciplines has been unable to reach a broad, popular audience (there is, for instance, very little linguistics in science museums). To remedy, and building on a previous project, we will create and test mixed physical/computational hands-on models of various linguistic phenomena, focusing like in T1 on language diversity and targeting schools and science exhibits.

Research Project Title

Technologies for Research and Education in Linguistics (TREiL)

2 State of the art

Theoretical linguistics (broadly understood, here, as a discipline spanning generative grammar, formal semantics, linguistic typology, with focus on diachronic or synchronic phenomena) is one of the few scientific fields which have never used technology as a core part of its methodology.

The present project starts from the perception of this gap. Its main claim is that there are various simple technological tools, borrowed from web design, computational linguistics and the tradition of hands-on models used for the popularization of 'hard sciences', which can improve the end-to-end process of doing linguistics, from data collection, through data analysis and interpretation, down to outreach dissemination of the results.

TREiL builds on two European projects its proponents are currently in: the ERC proposal COMPOSES (PI Marco Baroni; Co-PI Bernardi and Zamparelli; ending in 2016) and AThEME (an 18 partner "Cooperation" project on multilingualism; PI: Lisa Cheng). We will extend techniques developed in these projects, shifting the focus on the methodology of doing linguistics. While our project description makes reference to a broad 'generative' framework, we believe that our results could be applied to many approaches to human language.

The claim that theoretical linguistics does not use technology needs a justification, which also helps to pinpoint the aims of the proposed project. There exists, of course, a whole subfield of linguistics, *computational linguistics*, which relies on computers and machine-readable corpora, Its goals, however, are typically quite different from those of its theoretical counterpart. The latter aims at factoring the largest number of linguistic facts (within and across languages) into a smaller set of core principles and core language difference ("parameters"), making predictions on the range of variation that can be found in the world's languages.

The part of computational linguistics which aims to clarify human psychology (as opposed to developing customer-ready language tools) has fared best at modeling *graded* phenomena within a single language. Examples are the probability of assigning a certain parse to an ambiguous structure (Manning and Schütze, 1999, Nivre, 2003, Clark and Curran, 2007), resolving a lexical ambiguity (Yarowsky, 1995, Pantel and Lin, 2002, Navigli, 2009), or predicting the extent to which two words prime each other, two sentences are judged similar, (Dumais and Landauer, 1997, Padó and Lapata, 2007, Padó et al., 2007), or in an entailment relation (Kotlerman et al., 2010). Recent *distribution-based* approaches to computational semantics can even generalize to unseen combinations, predicting for instance how much unattested adjective-noun pairs will be considered semantically deviant by humans (Vecchi et al., 2015).

Despite all this, questions about the deep reasons behind the universal properties of languages, such as why languages which allow null subjects allow better subject extraction, why negative polarity items like *ever* occur only in certain environments; why no language agrees in gender but not in number, etc. have not been addressed with computational means.

There is, however, a growing interest in part of the computational linguistic community to seek explanatory accounts of the data observed, finding bridges with its theoretical counterpart (see the recent journal "Corpus Linguistics and Linguistic Theory"), reconciling for instance Montague-style compositional semantics with distribution-based models of lexical meanings (Baroni et al., 2014).

It seems to us, however, that there isn't at present a sufficient corresponding effort on the theoretical linguistic side to expand its toolbox and develop new methodologies for the collection, analysis and presentation of the data. Some linguists have embraced the use of statistics and p. values to analyze complex judgment patterns (Bard et al., 1996, Cowart, 1997, Sorace and Keller, 2005), and there is a growing interest in developing new tools for the process of data gathering (see e.g. the use of gaming in Poesio et al. 2013), but also in increasing the role of linguistics in education (see Hudson 2004, Denham and Lobeck 2014, and the journal "*Language and Linguistic Compass*"). Still, as a whole, theoretical linguistics is still largely free of technologies more complex than word processors, tape recorders, date bases and web queries. With TREiL, we hope to show that great advances can be made by using software tools which are already available, coupled with variations on models which have been popular for decades in physics and other hard sciences to improve the outreach of the discipline in a wider audience (e.g. schools, the visitors of science museums).

A PDF VERSION OF THIS PROPOSAL, WITH FULL REFERENCES, IS AVAILABLE AT THE URL: http://clic.cimec.unitn.it/roberto/PRIN2016-TREiL.pdf

3 Detailed description of the project: methodology, targets and results that the project aims to achieve and their significance in terms of advancement of knowledge

The process of doing theoretical linguistics can be broken up in three fundamental steps: gathering linguistic data; using the data to generate and test hypotheses against the background of previous research; writing up the results and making them available to the public. Our project has identified aspects within each of these steps which can be enhanced by employing a set of inexpensive and easily available technological tools. More specifically, the project aims to address the following problems.

1. DATA COLLECTION: LIMITS OF WEB DATA COLLECTION: Novel data collection is normally done in person (field work) or by questionnaires. In field work researchers can interact with speakers in various way (free conversation, guided conversation, interviews, etc.) and they know the context where the data are produced. However, the process is very labor-intensive, or not sufficiently precise: the judgments collected may be too few; there may be no certainty that the informants belong to the same linguistic community, or that they have filtered out irrelevant readings. Moreover, the informants might be insufficiently motivated to think through the examples or understanding the exact question, etc.

   The web can help in administering linguistic questionnaires to a large number of speakers, potentially yielding enough data to investigate very complex and subtle linguistic phenomena. However once again, (i) the informants must be motivated to access a certain web site and apply themselves to the task; (ii) the web does not lend itself to languages which are not normally written, like most of the world's dialects.

2. DATA ANALYSIS: HOW INNATE IS LANGUAGE?

   One of the crucial issues in the debate between generative linguistics in the Chomskian paradigm and empiricist approaches to language is the extent to which there is an innate *language faculty*, i.e. a specific language acquisition device which biases first language learners toward certain constructions, excluding others which (though functionally plausible) are never found in natural languages (Chomsky, 1986, Chomsky and Lasnik, 1993, Baker, 2001, Newmeyer, 2005). A 'pure' innatist would explain language universals in terms of underlying brain biases, a 'pure' empiricists, in terms of functional biases and statistical properties of the input, coupled with a general-purpose learning device (Clark and Lappin, 2010).

   In principle, one way to probe the potential role of language-specific innate biases would be to contrast the linguistic competence of a normal language learner with the competence of a learner who has acquired language without any pre-wired bias toward a specific analysis. Contrasting the 'biased' and the 'unbiased' learner on a number of linguistic tasks, and with a large number of data points, would clearly tell us a lot about which aspects of language, if any, are affected by innate knowledge.

   Unfortunately, such an experiment is impossible: we do not know where innate grammatical biases might be found in the brain, and even if we did we could not influence them, for

obvious ethical reasons.

3. OUTREACH: HOW TO GO BEYOND SPECIALISTS?

   Linguistic results appear in specialized publications, but unlike physics or other 'hard sciences', they have very little impact on the general public, if we set aside publications on how people ought to speak or write or word etymology. Books like Pinker (1994) or Moro (2015), which aim for the general public and try to present interesting facts about the structure of human language in an engaging, accessible style, remain exceptions.

   This leads to a remarkable degree of public ignorance about basic linguist facts. For instance, many educated people believe that the world has only one sign language, if they believe it is a language at all. In addition, when linguistic facts are taught in schools, they are handed down as rules plus exceptions to memorize. The notion that any speaker could carry out 'experiments' on language, that many exceptions might actually reveal hidden regularities, that probing language structures can be as entertaining as solving a puzzle, comes as a complete surprise to the students of an introductory linguistic course at university.

   A clear symptom of this state of affairs is that the formal side of language is virtually absent from science museum (even one of the few exceptions, the Language Science Research Lab at COSI, Wagner et al. 2015, does not seem to cover syntax). One reason is that linguistics does not even seem to have the (basic) technology to build the type of hands-on experiments that are central in today's science museums experience (Resnick, 2003).

Our claim is that an appropriate use of technology can make great advances on each of these problems.

3.1 Task 1. Data collection: expanding the VinKo project

The starting point for this task in our methodological project is the online platform VinKo, "Varietäten in Kontakt", designed by a group of researchers and technicians from the universities of Trento and Verona as part of the EU Cooperation project "AThEME", on multilingualism, and about to open to the public.

The purpose of VinKo is to gather a large amount of oral linguistic data (both sentences and lists of words) in all local Romance and German languages (minority languages and dialects) which are spoken in the area between Innsbruck and Verona: Trentino Romance dialects, Tyrolean dialects, Ladin, Cimbrian, Mòcheno. The web is used to distribute a questionnaire which (a) can be interrupted and resumed at a later time; (b) contains model sentences in Italian for Trentino and Ladino speakers; in German, for Cimbrian, Mòcheno and South-Tyrolean speakers; (c) proposes sentences which are simple and used normally in daily speech; (d) avoids priming effects by placing sentences appropriately. The researchers can filter the data thus collected using several criteria (variety, location, grammatical phenomenon, age of the speaker, etc.). The amount of data on specific grammatical phenomena which the system should be able to collect will give researchers the possibility of examining micro-variation in-depth, considering many types of variables at once.

The possibility of collecting oral data represents VinKo's most innovative feature. As mentioned in the list of problems above (1.ii), oral communication is particularly crucial for non-standard varieties, given the speakers' lack of familiarity with writing (Bel and Gasquet-Cyrus, 2011, 2013), and it is also the only way to gather large amounts of data on *prosody* in a semi-controlled environment. However, problem (1.i) still stands: how to motivate the speakers to use the system?

Our key idea is that the next releases of VinKo should be designed to increase the speakers' awareness that they are part of a linguistic community, triggering the interest in obtaining a detailed map of 'people who speak like me'—a particularly important marker of identity in a multilingual region like Trentino Alto Adige.

VinKo already shows a map where markers gradually appear to illustrate the geographical origin of the data recorded, but the new types of visualization we will focus on in TREiL will mark all the points where identical grammatical structures are attested, giving the speaker immediate feedback on his or her position on a number of isoglosses, from which—when this does not violate privacy—it will be possible to access the oral records of other speakers.

Building on Dorian (2014), Eisenlohr (2004), Grenoble and Whaley (2006,), we believe that this could greatly increase the motivation of speakers of local varieties to strengthen the use of the community language, spreading language knowledge and, finally, bringing youngsters in contact with elderly people.

VinKo has been so far funded as a part of the first phase of the AThEME grant. The AThEME budget will have no additional funds for future development of the tool, so securing an Italian grant would be instrumental in expanding the project as described.

3.2 Task 2. Data analysis: comparing natural and artificial grammaticality judgment

In the language innateness debate, corpus-based computational linguistics has traditionally sided with empiricism. There have in fact being several attempts to create general learning devices which, exposed to large amounts of language, could learn to perform tasks that seemingly require a linguistic competence (e.g. Question/Answering tasks, Lehnert 1977, Webber and Webb 2010, image caption generation, Karpathy and Li 2014, Xu et al. 2015, image-based Q/A Andreas et al. 2015, etc.). Many of these models use recurrent neural networks (RNN, Elman 1990), learning devices which take in input a string of symbols (e.g. words) and produce a useful output (e.g. a prediction of the probability of the following words); since the output is a function of the current input and the memory of previous ones, RNN can keep track of sequences in time. Several recent advances in RNN (see a bibliography in Choi and Kim 2015), especially LSTM networks (Hochreiter and Schmidhuber, 1997) and work building on them (Mikolov 2014, Graves et al. 2014, a.o.), coupled with 'deep' NN structures (LeCun et al., 2015), have been quite successful at modeling language, making headway even on long distance dependencies (Wh-questions, relative clauses, etc. see e.g. Olah 2015). Note that such models are not biased toward any particular language feature or construction; given sufficient memory, they should be able to learn equally well artificial languages with structures that do not exist in any natural language. Thus, they are the perfect embodiment of a 'pure' empiricist language device, at least on the syntactic side.

The goal of this TREiL Task is to systematically compare judgments given by humans with 'judgments' extracted from these RNN models of language. Of particular interest will be the status of pairs of constructions which have very low frequency (so require some kind of generalization from the actual learning material), yet diverge in grammaticality: extraction islands (from coordination (1), subject (2), etc.), long-distance selectional violations ("He had *murdered* something which turned out to be a *policeman/??number*"), but also agreement mismatches and other cases.

(1) a. A text that John wrote and Sue revised
b. ??A text that John wrote and Sue revised the footnotes.

(2) a. What do you think Marc has evidence for?
b. *What do you think evidence for nails Marc?

If human judgments on such data could be shown to be more clear-cut than any statistics-based prediction, this could be evidence for an innate component. Moreover, exploring a wide range of violations will give us the theoretical possibility of distinguishing language aspects or areas which are more 'empirical' from others that are more 'innate', giving a more shaded (and possibly more correct) view of the empiricism/innatism dilemma. To this effect, we will compare judgments on constructions which are impossible in the training language (English or Italian) but exist in other languages (e.g. Verb second) with judgments on constructions which are known not to exist universally (e.g. Verb penultimate).

3.3 Task 3. Outreach: bringing theoretical linguistics to a broader audience

In another part of the AThEME project, Roberto Zamparelli and a post-docs have worked on a novel approach to increase the outreach of theoretical linguistics: designing a physical, structural model of sentence, which can be assembled by a group of players as a kind of educational game. The words in the model connect in agreement to current syntactic theories, and they lock together only when they form syntactically correct sentences.

The key idea behind this experiment is to extend to formal linguistics the hands-on approach used in science museums (e.g. San Francisco's *Exploratorium*) to teach physics, chemistry and other disciplines. It is now well documented that hands-on, "project based" learning (Boss and Krauss, 2007) can greatly increase student's motivation in schools, and lead to an exploratory attitude toward the discipline, something which linguistics sorely misses. Indeed, while there are many games which make language their focus, they are all restricted to the lexical level (like e.g. *Scrabble*), and do not rest on a solid theoretical underpinning. Since the model is multilingual (currently, English, Italian and German), the players become immediately aware of the structural difference between the three languages and reuse this knowledge as they learn them as foreign languages.

The current models is almost ready to be tested in schools, but it is of course very far from being representative of the main themes in linguistics. The TREiL project would give us the opportunity to create new models covering morphology (here the building blocks would be morphemes, not words) and if possible phonology—the possibilities to explore are endless.

Moreover, while we believe that using physical pieces gives us an edge over purely computer based tools, since it gives people the possibility of manipulating real objects, we believe that the model could be enhanced by adding a computational side, interfaced via a smart-phone. In this expansion, the model pieces would carry a bar code, which, once scanned with a phone camera, could give the players additional info, project images of the object whose description is being assembled, examples of its use in different languages, etc.

Like with VinKo, the "language jigsaw puzzle" project has been funded as part of an initial phase of the AThEME grant, and will not receive further funding. Funds from TREiL would cover a follow up testing phase in schools, subsequent model tuning, the inclusion of a bar code and the development of cell-phone-based apps to read their information, as well as for the study of extensions to other aspects of linguistics (diachronic lexical change, morphological composition, possibly phonology).

4 Project development, with identification of the role of each research unit and research organizations involved, with regards to expected targets, and related modalities of integration and collaboration

Since the project is based at the University of Trento, across the Department of Humanities ("Dipartimento di Lettere e Filosofia") and the Center for Mind and Brain Sciences (CIMEC) there is only one research unit.

The overarching team in TREiL is the use of technology (sometimes cutting-edge software like LSTM networks, sometimes 'low tech' wood or plastic modules with bar codes) to achieve three important advances in theoretical linguistics: better distributed data gathering, a theoretically significant evaluation of artificial language models, and new, more entertaining ways of making a wide audience aware of linguistic results. These aspects are not only connected by a methodology: Task 2 requires a large amount of diverse human judgments, which can be gathered with the VinKo web site. In turn the models developed for Task 3 are (crucially) multilingual and education to a multilingual environment is crucial to the philosophy behind VinKo, in Task 1.

Still, the tasks are sufficiently independent to be partly pursued in parallel.

Task 1 could start immediately after the necessary technical personnel has been hired (see next section). After the current VinKo web side is opened to the public its pattern of usage will be observed for a few months (Task 1.1) and the quality of the data will be probed (Task 1.2)

After this evaluation phrase, we will discuss the design of the mapping interface, and experiment with it (Task 1.3). The approach chosen will be implemented and tested (Task 1.4). The revised web site will go online and will be publicized via social networks and through the culture institutes in the area (see next section). Next, we will analyze the patterns of usage of the new site, compare the differences and make any required modification (Task 1.5).

Following this phase, we will study the possibility to use VinKo to acquire intonational data (Task 1.6), by giving participants a text to read in different situations, and verify to what extent the recording can be mapped onto common intonational pattern. During Task 1.5 we will run VinKo

on Italian or English speakers (which will involve translating the web site) with a questionnaire containing the grammaticality judgments for Task 2.

Task 2 builds on the expertise acquired by two of the project members (Raffaella Bernardi and Roberto Zamparelli) in the ERC COMPOSES grant, ending in 2016. COMPOSES used RNNs to generate various data sets (e.g. the forthcoming LAMBADA data-set), and provided the hardware (clusters and GPU-equipped PC) and the related technical expertise.

This part of the research will take place at CIMEC (http: http://web.unitn.it/cimec), the Interdepartmental Center on Mind and Brain sciences of the University of Trento. With its CLIC Lab on language and computation, CIMEC, a renowned international research center with a very interdisciplinary approach, will provide the ideal setting for this part of the project, including all the technical environment to carry out comparisons with ERP data.

In a first phase of of this part of the project (Task 2.1) we will select and train appropriate RNN models, then explore ways to obtain the closest equivalent of a grammaticality judgment from the trained model by looking at their activation patterns after it receives ungrammatical sentences (e.g. by means of a classifier). We will then select from the theoretical literature a variety of semantic and syntactic violation which could suitable for our comparison and obtain qualitative and quantitative judgments from human speakers on these violations (Task 2.2). After a pilot study, we will make use of the VinKo web site developed in Task 1 to obtain an amount of judgment data adequate for statistical analysis. This time we will target speakers of standard Italian, German or English, and possibly fill out missing data points using the crowd-sourcing methods we used in e.g. Vecchi et al. (2015), Marelli et al. (2014) (Task 2.3). We will then compare the two sets of results, adjusting the network parameters (especially, memory and training regime) to minimize their distance. Any residual difference will be analyzed analytically (Task 2.4). Finally, we will compare the results we obtain from the model with the literature on ERP patterns in the presence of the same type of violations (Friederici et al., 1993, Kutas and Federmeier, 2011), to see if the way the RNN reacts to different types of ungrammatically is comparable to brain data for similar violations (Task 2.5).

By the time the project starts, Task 3 should be able to count on a working prototype of the "Language Jigsaw Puzzle" representing the syntax of German, English and Italian. After the hiring phase (see next section), we will begin to work on three tasks: usability improvements for the hands-on language tools we will have created (Task 3.1); expanding the hands-on method to morphology and phonology (Task 3.2); using bar codes and a phone code-scanning app to add information to the word or morpheme pieces, so that the users can have semantic information about the pieces that he or she is trying to combine. To this effect we will build on distributional semantics, specifically on the metrics of word similarity and word compatibility (Vecchi et al., 2015) developed within the COMPOSES project, but extended to multiple languages. This will give the use the possibility of automatically associating a part of the structure which is being built (say, the VP "was a red herring") with both syntactic and semantic information, such as phrases with similar or identical meaning in other languages ("era una falsa traccia"), pictures of actual herrings, words similar to "herrings" and to "red", etc.

We have already established contacts with MUSE (http://www.muse.it/) , an important science

museum in Trento (specifically, the Muse FAB LAB is helping us build our language puzzle prototypes) and we plan to try to test, and eventually, display the models we create in one of the museum's exhibits.

4.1 Contribution of individual team members

The following people are expected to work in TREiL:

- Roberto Zamparelli (CIMEC, Dept. of Psychology and Cognitive Science), associate professor, PI;
- Raffaella Bernardi (CIMEC, Dept. of Informatics), researcher;
- Patrizia Cordin (Dept. of Humanities), associate professor;
- Manuela Moroni (Dept. of Humanities), researcher

In addition, we intend to hire two post-docs: one for Task 2, one for Task 3 of the project, and a technical collaborator, to work on the VinKo web site development.

While the three tasks in which the project is divided might seem quite broad, they are actually based on proven expertise within UNITN, and they all builds on previous work of ours. At the same time, each of the tasks breaks new ground on a different front: a novel method for data acquisition, a novel approach for comparing traditional linguistic data with cutting-edge research in computational linguistics, and a new and much needed effort to show to a broader audience unexpected sides of language research, potentially offering new tools for language learning.

Our curricula fit perfectly with these goals. Patrizia Cordin has an extensive record of work on dialectal variation in the Trentino area, and she will lead the development of Task 1, after co-supervising the design of the original VinKo web site.

Manuela Moroni, an expert in prosody, will be consulted to devise the best way to collect intonational profiles in Task 1, a possibility given by VinKo's emphasis on oral data.

The role of the technical collaborator is quite central to the project. The web site will require strict monitoring for at least 2 months after it is open to the public, to check if the interface works and if the data base copes with the (potentially large) mass of oral data. Further modification must be carried out by an expert in web design, since the interface should remain sufficiently polished. None of the faculty has the expertise or time to do this, so external collaboration is essential.

The post doc we will hire for Task 2 should be an expert in computational linguistics, with a good experience with recent RNN models and ideally some working knowledge of theoretical linguistics. He or she should contribute to all aspects of Task 2.

The post doc to hire for Task 3 should be very familiar with CAD software, and also capable of develop apps for the Android or iOS operating systems. Italian proficiency is required, as he or she should interact with school environments during testing.

Turning to the CIMEC tenured faculty, Raffaella Bernardi is an expert in computational linguistics, formal and distributional semantics, and has been one of the leader in the 2014 LREC

exercise, based on the SICK paraphrase data-set we developed (Marelli et al., 2014). She will contribute to Task 2. With Roberto Zamparelli, she has worked on bridging the gap between the formal linguistic view of compositionality and distributional semantics (Baroni et al., 2014). Task 2 has the same aim, just a different methodology.

Roberto Zamparelli has worked on the syntax-semantics interface and on lexical semantics, but also on computational semantics within COMPOSES (see especially Baroni and Zamparelli 2010). He has designed and built the very first (monolingual) prototype of the "Language Jigsaw Puzzle" for the 2011 "Notte dei Ricercatori". As the UNITN coordinator for the European Collaboration project "AThEME", he has expertise in directing a research group, and has been part of several other research projects (see the CV).

5 Possibile application potentialities and scientific and/or technological and/or social and/or economic impact of the project

Since TREiL focuses on how to use technologies in order to foster a close interaction between researchers and the general public (speakers of endangered languages, non-profit societies involved in preservation and revitalization, educators, cultural operators, students), we expect a lively cultural and social impact. In particular we expect that:

(i) TREiL should consolidate the collaboration between national and international team networks, extending it to new partnerships. In this perspective, the possibility of collaborating with academic institutions interested in the systematic collection of data and information on dialects and endangered languages becomes a central requirement.

The data acquired with VinKo should be made freely available via an online research infrastructure accessible to all interested stakeholders, with the possibility of searching, organizing, visualizing and analyzing all non-sensible data. We will consider whether oral data might violate privacy, bring up the matter with the local ethical committee and take appropriate measure, including releasing some data only as aggregates.

(ii) TREiL would consolidate the collaboration with the regional linguistic minorities cultural institutes: the Ladin Cultural Institutes (Majon di Fascegn and Micurà de Rä), the Mòcheno Cultural Institute (Bersntoler Institut), the Cimbrian Cultural Institute (Kulturinstitut Lusern);

(iii) events (workshops and seminars) will be organized for both academic and non-academic audiences (the Provincial Service for the promotion of linguistic minorities, the Ladin Cultural Institutes, the Mòcheno Cultural Institute, the Cimbrian Cultural Institute, Olfed (Ofize Ladin Formazion e Enrescida Didatica) and Sorastanza of Fassa School);

(iv) lectures and demos for schools will be organized in order to present the data that have been gathered and discuss them with teachers and students;

(v) once Task 3 has obtained and tested hands-on prototypes for various aspects of language, we will bring them to schools and local science exhibits, and we will consider industrial production.

On the scientific side, the comparison between artificial language models, trained on statistical properties of the input and the competence of a human being, as evidenced by grammaticality judgments, is a novel methodological approach which could bring the two sides of linguistic science much closer, and speak, potentially, to both research communities.

References

J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. *CoRR*, abs/1511.02799, 2015. URL `http://arxiv.org/abs/1511.02799`.

M. C. Baker. *The atoms of language: the mind's hidden rules of grammar.* Oxford University Press, Oxford, 2001.

E. Bard, D. Robertson, and A. Sorace. Magnitude Estimation of linguistic acceptablity. *Language*, 72: 32–68, 1996.

M. Baroni and R. Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA, 2010.

M. Baroni, R. Bernardi, and R. Zamparelli. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies*, 9(6):5–110, 2014.

B. Bel and M. Gasquet-Cyrus. Interdisciplinarity and the sharing of oral data open new perspectives to field linguistics. In *Colloque de l'AFLS: Regards nouveaux sur les liens entre théories, méthodes et données en linguistique française*, Nancy, France, 2011. URL `http://lpl-aix.fr/ fulltext/4733.pdf`.

B. Bel and M. Gasquet-Cyrus. Digital curation at the service of endangered languages. In M. Jones and C. Connolly, editors, *Endangered Languages and New Technologies*. Cambridge University Press, Cambridge, 2013.

S. Boss and J. Krauss. *Reinventing Project-Based Learning: Your Field Guide to Real-World Projects in the Digital Age.* ERIC, 2007.

M. Choi and J. Kim. Awesome recurrent neural networks. A curated list of resources dedicated to recurrent neural networks; GitHub, 2015. URL `https://github.com/kjw0612/awesome-rnn`.

N. Chomsky. *Knowledge of Language*. Praeger Publications, New York, 1986.

N. Chomsky and H. Lasnik. The theory of principles and parameters. In J. Jacobs and al., editors, *Syntax: An International Handbook of Contemporary Research*, volume 1, pages 506–569. Walter de Gruyter, 1993.

A. Clark and S. Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley Blackwell, 2010.

S. Clark and J. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

W. Cowart. *Experimental syntax: Applying objective methods to sentence judgments*. Sage Publications, Thousand Oaks, 1997.

K. Denham and A. Lobeck, editors. *Linguistics at School: Language Awareness in Primary and Secondary Education*. Cambridge University Press, 2014.

N. Dorian. *Small-Language Fates and Prospects: Lessons of Persistence and Change from Endangered Languages: Collected Essays*. Brill, Leiden, 2014.

S. Dumais and T. Landauer. A solution to Platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological review*, 104:211–240, 1997.

P. Eisenlohr. Language revitalization and new technologies: Cultures of electronic mediation and the refiguring of communities. *Annual Review of Anthropology*, 33:21–45, 2004. Suggerito da Cordin.

J. Elman. Finding structure in time. *Cognitive Science*, 14:179–211., 1990.

A. D. Friederici, E. Pfeifer, and A. Hahne. Event-related brain potentials during natural speech processing: effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, 1 (3):183 – 192, 1993. ISSN 0926-6410. doi: http://dx.doi.org/10.1016/0926-6410(93)90026-2. URL http://www.sciencedirect.com/science/article/pii/0926641093900262.

A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. http://arxiv.org/abs/1410.5401, 2014.

L. A. Grenoble and L. J. Whaley. *Saving Languages: An Introduction to Language Revitalization*. Cambridge University, New York, 2006,.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. URL http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf.

R. Hudson. Why education needs linguistics (and vice versa). *Journal of Linguistics*, 40:105–130, 2004.

A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014. URL http://arxiv.org/abs/1412.2306.

L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 2010.

M. Kutas and K. D. Federmeier. Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (erp). *Annual Review of Psychology*, 62:621–647, 2011. doi: http://doi.org/10.1146/annurev.psych.093008.131123.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

W. Lehnert. Human and computational question answering. *Cognitive Science*, 1 (1):47–73, 1977. ISSN 1551-6709. doi: 10.1207/s15516709cog0101$_3$. URL http://dx.doi.org/10.1207/s15516709cog0101$_3$.

C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambrdige, MA, 1999.

M. Marelli, S. Menini, M. Baroni, R. Bernardi, and R. Zamparelli. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*, 2014.

T. Mikolov. Using neural networks for modeling and representing natural languages. In *COLING 2014, 25th International Conference on Computational Linguistics, Tutorial Abstracts, August 23-29, 2014, Dublin, Ireland*, pages 3–4, 2014. URL http://aclweb.org/anthology/C/C14/C14-3002.pdf.

A. Moro. *I confini di Babele. Il cervello e il mistero delle lingue impossibili*. Il Mulino, 2015.

R. Navigli. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

F. J. Newmeyer. *Possible and Probable Languages. A generative Perspective on Linguistic Typology*. Oxford University Press, 2005.

J. Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of IWPT*, pages 149–160, Nancy, France, 2003.

C. Olah. colah's blog: Understanding-lstms. 2015. URL http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

S. Padó and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

U. Padó, S. Padó, and K. Erk. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP*, pages 400–409, Prague, Czech Republic, 2007.

P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of KDD*, pages 613–619, Edmonton, Canada, 2002.

S. Pinker. *The language instinct*. William Morrow, New York, 1994.

M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, Apr. 2013. ISSN 2160-6455. doi: 10.1145/2448116.2448119. URL http://doi.acm.org/10.1145/2448116.2448119.

M. Resnick. Playful learning and creative societies. *Education Update*, 8(6), 2003.

A. Sorace and F. Keller. Gradience in linguistic data. *Lingua*, 115:1497–1524, 2005.

E. Vecchi, M. Marelli, R. Zamparelli, and M. Baroni. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, 2015.

L. Wagner, S. R. Speer, L. C. Moore, E. A. McCullough, K. Ito, C. G. Clopper, and K. Campbell-Kibler. Linguistics in a science museum: Integrating research, teaching, and outreach at the language sciences research lab. *Language and Linguistics Compass*, 9(7):420–431, 2015.

B. Webber and N. Webb. *Question Answering*, pages 630–654. Wiley-Blackwell, 2010. ISBN 9781444324044. doi: 10.1002/9781444324044.ch22. URL http://dx.doi.org/10.1002/9781444324044.ch22.

K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. URL http://arxiv.org/abs/1502.03044.

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics. doi: 10.3115/981658.981684. URL http://dx.doi.org/10.3115/981658.981684.