ERC Advanced Grant 2015
Research proposal [Part B1]

Strengthening Language Technology with Linguistic Theory

WEIGHT-TRAINING

Principal Investigator: **Laura A. Janda**
Host Institution: **UiT, The Arctic University of Norway**
Proposal Duration: **60 months**

Proposal summary:

WEIGHT-TRAINING applies a novel methodology to theoretical and practical questions of language complexity by implementing the use of weights in computational models of natural language morphology.

In most languages, words can have multiple forms, comprising morphological paradigms. Paradigms can be complex and present challenges for both linguistic analysis and language technology. If we are to come to grips with morphological complexity, Russian is an optimal starting point because it maximizes the combination of morphological complexity with a deep tradition of linguistic analysis and vast data resources.

WEIGHT-TRAINING will build the first full-scale weighted computational model of Russian morphology and extend the weighted model to other languages, including complex and minority circumpolar languages. This extension will demonstrate that the model is applicable to other languages and put a powerful tool in the hands of linguists with far-reaching implications for linguistic theory.

In the late 20th century it was assumed that differential assessment of language complexity was motivated by racist attitudes. Recently this assumption has been debunked, and there is evidence that languages with fewer speakers tend to exhibit greater morphological complexity. Many metrics of morphological complexity are reductionist by necessity, as in typological surveys that must code differences across a vast number of languages.

WEIGHT-TRAINING is by contrast an in-depth study of morphological complexity in a select group of languages representing various morphological types. WEIGHT-TRAINING combines the strengths of linguistic theory with language technology to challenge assumptions about the structure of morphological paradigms. WEIGHT-TRAINING addresses a serious blind spot by focusing on morphologically complex and minority languages, which tend to receive less attention from non-typological theoretical linguists and to be under-resourced in terms of technology.

Section a: Extended Synopsis

WEIGHT-TRAINING: Strengthening Language Technology with Linguistic Theory
WEIGHT-TRAINING applies a novel methodology to theoretical and practical questions of language complexity in the way it implements weights in computational models of natural language morphology.

In most languages, words can have multiple forms, comprising morphological paradigms. Paradigms can be complex and present challenges for both linguistic analysis and language technology. If we are to come to grips with the complexity of morphological paradigms, Russian is an optimal starting point because Russian maximizes the combination of morphological complexity with a deep tradition of linguistic analysis and vast data resources (in particular, collections of texts known as corpora).

WEIGHT-TRAINING will build the first full-scale open-source weighted computational model of Russian morphology. WEIGHT-TRAINING will then extend the weighted model to other languages, including complex and minority circumpolar languages. Thus we will demonstrate that the model is extendable and that it has far-reaching implications for linguistic theory.

In the late 20th century it was assumed that language complexity was related to "primitiveness", and that differential assessment of complexity was motivated by racist attitudes. In the past decade this assumption has been debunked (Joseph & Newmeyer 2012), and there is growing evidence that languages with fewer speakers tend to exhibit greater complexity, particularly in terms of morphology (Trudgill 2011, McWhorter 2011). Many metrics of morphological complexity are reductionist either by necessity, as in Lupyan & Dale's (2010) survey of over 2000 languages, or by design, like Corbett's (2015) survey of splits in morphological paradigms: both evaluate phenomena in terms of +/- values.

WEIGHT-TRAINING is by contrast an in-depth study of morphological complexity in a select group of languages. WEIGHT-TRAINING combines the strengths of linguistic theory with language technology to challenge assumptions about the structure of morphological paradigms. WEIGHT-TRAINING addresses a serious blind spot by focusing on morphologically complex and minority languages, which tend to receive less attention from most theoretical linguists and to be under-resourced in terms of technology.

WEIGHT-TRAINING will impact linguistic theory by mapping out the realistic size and structure of paradigms and compare their morphological complexity with theoretical models and measures at a level of detail that has not been achieved for our selected languages. We will build morphological analyzers that can extract meaningful quantitative data for future analysis and are designed to be extendable to other languages, giving a significant boost to empirical approaches to linguistics. Computational linguistics, largely dominated by stochastic models, will benefit by our facilitation of a theoretical direction for this line of research. Our weighted models for morphology will yield better electronic resources for language users and learners. Our focus on minority and complex languages will promote pluralistic and equitable societies in the circumpolar region.

WEIGHT-TRAINING unfolds through the seven objectives visualized in Figure 1 and described below.

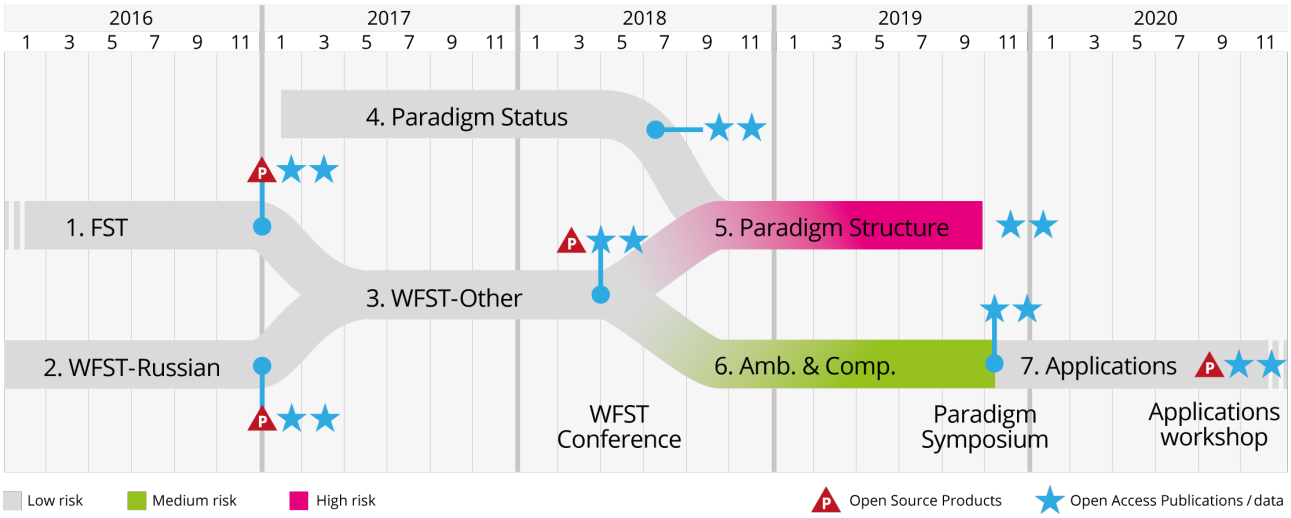


Figure 1: WEIGHT-TRAINING Gantt Chart

Objective 1. "FST": Complete finite state transducers for selected circumpolar and other languages. Finite state transducers (FSTs; Roche & Schabes 1997) are the *de facto* standard for modeling natural language morphology. FSTs are computationally efficient, compact, and reversible: the same model performs both analysis and generation of wordforms. FST performance is sufficient for morphological

analysis at 90% and ceilings at 98%. Table 1 shows a sample of FSTs available at UiT's Giellatekno, the Center for Saami language technology [1] to be included in WEIGHT-TRAINING. Five FSTs already have ≥90% coverage, and the remainder are already above 50%, so we have a large head start on this objective.

Language	Family : Group	Type	Number of Speakers	FST Coverage
Russian	Indo-European : Slavic	Fusional	166,167,860	97%
Norwegian Bokmål	Indo-European: Germanic	Fusional	4,640,000	96%
Swedish	Indo-European: Germanic	Fusional	9,200,000	90%
Icelandic	Indo-European: Germanic	Fusional	313,840	88%
Finnish	Uralic : Finnic	Agglutinating	5,100,000	98%
Kven Finnish	Uralic : Finnic	Agglutinating	2,000	51%
North Saami	Uralic : Saamic	Fus/Agg	25,700	98%
Inari Saami	Uralic : Saamic	Fus/Agg	300	53%
Nenets	Uralic : Samoyedic	Agglutinating	22,000	56%
Erzya	Uralic : Mordvinic	Agglutinating	260,000	89%
Udmurt	Uralic : Permic	Agglutinating	339,800	71%
Komi-Zyrian	Uralic : Permic	Agglutinating	156,000	76%
Evenki	Tungusic: Evenki	Agglutinating	17,000	51%
Yakut	Turkic: Northern Turkic	Agglutinating	360,000	70%
Greenlandic	Aleut-Inuit-Yupik: Inuit	Polysynthetic	60,000	83%
Plains Cree	Algonquian : Cree	Poly/Fus	34,000	70%

Table 1: FST coverage for WEIGHT-TRAINING languages

The aim is to design a model that can be extended particularly to minority and complex languages. Shared features and contact relationships motivate the selection of Indo-European and Uralic languages; additional languages challenge our model and extend our circumpolar scope.

Objective 2. "WFST-Russian": Set weights in the Russian FST.
Is the paradigm just a list of equally probable items as commonly assumed (cf. McCarthy 2005)? Without weights, an FST mimics this assumption, but paradigm forms can differ greatly in frequency. Figure 2 shows the grammatical profile (cf. Janda & Lyashevskaya 2011) of Russian abstract nouns ending in *-ost'*. We see that some forms (nominative and genitive singular) are very common, while others (all plurals) are rare.

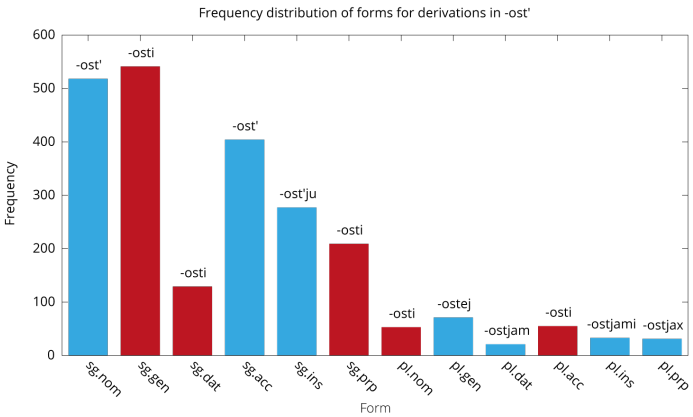


Figure 2: Grammatical profile of Russian abstract nouns in *-ost'*; intraparadigmatic homonyms in *-osti* marked in red (see objective 6).

WEIGHT-TRAINING will set weights to model the probability of morphological forms. For Russian, various kinds of information can be combined and contrasted as input for weight-setting algorithms: 1) frequencies attested in large corpora, 2) theoretical insights about classes of words and morphological categories, and 3) psycholinguistic experiments. While corpus frequencies (stochastic approach, cf. Droste et al. 2009) are valuable: a) they do not package information in a way that facilitates further insights; b) the majority of the world's languages are small (Lewis et al. 2015) and lack the resources to collect large corpora, so this option is available only to a few highly privileged languages with huge numbers of speakers; and c) small languages are the ones that tend to be morphologically complex (Lupyan & Dale 2010). The strategy of WEIGHT-TRAINING is to use Russian as a testing ground to determine which theoretically motivated weights are most robust vis-a-vis statistical distributions.

Objective 3. "WFST-Other": Set weights in FSTs for selected circumpolar and other languages.
Generalizations from the Russian WFST will be, with appropriate adjustments, extended to the other languages in Table 1. We will use statistical distributions of forms that are measurable in Russian to make educated guesses about languages where the opportunity to measure distributions is limited.

"FST-Other", "WFST-Russian", and "WFST-Other" will yield full-scale free and open-source language models for the selected languages. These WFST models put a powerful tool in the hands of linguists by making it possible to more reliably track the behavior of morphological features in any text. Scholarly results will be presented at our WFST Conference and published as open access articles.

Objective 4. "Paradigm Status": Map the status of the paradigm.
Maybe you don't know what a bittern is (it's a type of bird), but if I ask you for the plural, you will reply "bitterns" because you know how to form plurals in English. The paradigm of this word has two forms: *bittern, bitterns*.

Paradigms have been with us for more than 3,000 years, since the Old Babylonians, but there are reasons to doubt their psychological reality. Zipf's law (1949) describes the skewed distribution of words in a corpus, with a few words of high frequency, a sharp decline, and many words of low frequency. Since Zipf's law applies to wordforms, the number of words that appear in all their wordforms (complete paradigm) in a corpus is small, and this number quickly drops toward zero as the paradigm grows larger. In a pilot study we found that fewer than 30% of English nouns in the Baby BNC corpus (4M words) appear in both singular and plural forms. Norwegian nouns can express definiteness (roughly equivalent to *the* vs. *a*) in addition to singular and plural, and thus have a paradigm of four forms, but only 3% of nouns in the Norwegian Dependency Treebank (0.3M words) appear in all four forms. Russian nouns can have six cases expressed in singular and plural yielding a full paradigm of twelve forms, but only 1% of nouns appear in twelve forms in the Russian National Corpus Gold Standard (1.3M words; "gold standard" means hand annotated). Czech has seven cases and therefore fourteen forms in its noun paradigm, but no noun appears in more than twelve forms in the Prague Dependency Treebank (0.36M words). North Saami has 130 cells in its noun paradigm (Nickel & Sammallahti 2011), but a manual analysis of over 0.66M words (Antonsen & Janda forthc.) reveals that 36 of the paradigm forms are never attested for any noun in our corpus.

Since Zipf's law scales up, one could hypothesize that a speaker's total exposure to her/his native language is like a very large corpus with the same properties. This means that 70% of English nouns and 97% of Norwegian nouns will never be encountered in their full paradigms by native speakers of those languages. Native speakers of Czech and Russian will be exposed to full paradigms for 1% or fewer of their nouns. And a native speaker of North Saami might never encounter all the forms in the noun paradigm of that language at all, much less all forms of any single noun. So does the complete paradigm exist? Or is the paradigm just a convenient fiction for linguists? If so, what are the implications for linguistic analysis and computational modeling? WEIGHT-TRAINING aims to answer these questions.

This objective will focus on the five circumpolar languages for which we have access to gold standard corpora (see Table 2), plus North Saami, for which we will build a gold corpus of about 0.5M tokens. We will map the degree to which paradigm forms are attested with attention to factors such as part of speech, semantic class, and semantics of morphological categories.

Language	Gold Corpus Name	Tokens in Gold Corpus
Russian	RNC: Gold Standard	1,328,465
Norwegian	Norwegian Bokmål Dependency Treebank	311,277
Swedish	Talbanken	96,346
Icelandic	Icelandic Language Corpus (Mim)	1,031,209
Finnish	FinnTreeBank	208,101
North Saami	North Saami Gold Standard (projected)	500,000

Table 2: Languages in which we will map the status of the paradigm

Objective 5. "Paradigm Structure": Determine the structure of paradigms.
WFSTs are built upon the observation that the forms of a paradigm are not equiprobable. Do differences in the frequencies of paradigm forms reveal the structure of paradigms (cf. Bybee 1985, Karlsson 1985, Janda 2007)? Nesset & Janda (2010) have suggested that paradigms are structured and evolve as radial categories. In a radial category a network of related items is organized around a prototype (Janda 2006, 2010, 2015). WEIGHT-TRAINING tests the hypothesis that paradigms have a radial category structure and that this structure can inform a scheme for setting weights.

A paradigm split divides wordforms into two or more groups. For example, the Russian verb for 'drink up' is split into: *vypi*+endings for past tense and infinitive, *vyp'j*+endings for present tense, and *vypej*+endings for imperative. Corbett (2015) asks: What kinds of splits are possible? How are splits related to factors such as markedness (Janda 1995, Trosterud 2004)? We need this information in order to build reliable resources for languages. WEIGHT-TRAINING will answer these questions with reference to the languages in Table 1 and provide hypotheses for language evolution.

Objective 6. "Ambiguity & Complexity": Assess the relationship between ambiguity and complexity.
Ambiguity has proven a real hardship for corpus linguists. Russian abstract nouns like *radost'* 'joy' (Janda & Solovyev 2009) have a form *radosti* that is ambiguous, with five possible readings (see red columns in Figure 2). We call these forms "intraparadigmatic homonyms". In a pilot study we found that 35% of tokens in a Russian corpus exhibit this type of ambiguity. The Russian imperative form *vypej!* 'drink up!' (выпей! in Cyrillic, see Figure 3) is also the genitive/accusative plural form of *vyp'* 'bittern'; these are "morphosyntactically incongruent homonyms" and 10% of tokens in a Russian corpus exhibit this type of ambiguity. "Morphosyntactically congruent homonyms" are of the type *leču*, which can mean both 'I fly' and 'I treat (medically)' in Russian; these are less common (1% of tokens), but more onerous for disambiguation. In sum, over 45% of words in a Russian corpus are ambiguous.

Figure 3: Ambiguity in Russian: выпей! (*vypej!*) means both 'drink up!' and 'bitterns!' (genitive/accusative plural)

Ambiguity is by no means peculiar to Russian. For example, our FST-sourced electronic dictionary of North Saami [2] lists 16 interpretations for *nuorat* as various forms for 'young', 'youth' and 'make blunt'. WEIGHT-TRAINING will use WFST models to address disambiguation problems.

It is not known whether languages with greater morphological complexity also present larger ambiguity issues. For example, the North Saami noun paradigm with 130 cells has only 91 distinct forms; the remainder are at least two-way ambiguous (Janda & Antonsen under submission). WEIGHT-TRAINING will track the relationship between morphological complexity and ambiguity in detail for our focused set of languages (Table 1). The results of "Paradigm Status", "Paradigm Structure", and "Ambiguity & Complexity" will yield an open-access scholarly anthology and will be the topic of our Paradigm Symposium.

Objective 7. "Applications": Implement and test WFST models in applications for users and learners.
The weighted FST models built by WEIGHT-TRAINING will be implemented in a variety of electronic resources that service both native users and language learners, including machine translation (Unhammer & Trosterud 2009, 2012), electronic dictionaries (that look up a word in any of its forms; Tyers et al. 2009, Johnson et al. 2013), spellcheckers (Antonsen 2012), and ICALL modules (Intelligent Computer-Assisted Language Learning; Antonsen et al. 2009, Antonsen et al. 2013) that support language learning, maintenance, and revitalization. Because WEIGHT-TRAINING is specifically designed to be extendable to complex and minority languages, it removes existing barriers to access to such resources for these languages.

The world's linguistic legacy faces a situation that Harrison (2007) labels an "erosion of human knowledge": 90% of minority languages are expected to be replaced by dominant languages by the end of the 21st century (UNESCO 2003). We are losing precisely the small languages that are likely sources of data on morphological complexity, and in the Arctic this situation is acute due to climate change and the exploitation of mineral resources (Arctic Council [3]; Freeman 2000).
UNESCO ([4] 2014) recognizes that there are deep inequalities in access to language technologies and supports the inclusion of all languages in the digital world in order to foster "pluralistic, equitable, open and inclusive knowledge societies".

This activity culminates in the release of free open-source products, open-access scholarly works, and a Workshop highlighting comprehensive resources that support the vitality of languages.

Table 3 details the WEIGHT-TRAINING Team in terms of commitment, financing, and expertise.

Team member	%W-T : length in months	financing %ERC : %UiT	Linguistic expertise		Linguistic theory	FST expertise	FST theory	applications expertise
			Russian	Uralic				
Janda	50% : 60	50% : 50%	+++	++	+++	+		++
Trosterud	20% : 60	20% : 80%	++	+++	+++	+++	+	+++
Nesset	20% : 60	20% : 80%	+++		+++			++
Ylikoski	20% : 60	20% : 80%	+	+++	+++			+
Antonsen	20% : 60	20% : 80%		+++	++	+++		+++

Table 3: WEIGHT-TRAINING Team Composition (competence: + = some, ++ = moderate, +++ = strong)

Laura A. Janda, professor of Russian linguistics, is a recognized leader in cognitive linguistics (President of the International Cognitive Linguistics Association 2007–2011, Associate Editor of *Cognitive Linguistics* 2008-present). She has done pioneering work on morphological categories of case and aspect in Slavic languages and on North Saami ambipositions and possessive suffixes. Together with members of the CLEAR (Cognitive Linguistics: Empirical Approaches to Russian) research group at UiT, Janda has elaborated a suite of quantitative methods known as linguistic profiling. Her scholarly work has been published in over 100 articles and 18 books, and she has authored multimedia materials for language learners. She has led or co-led 9 major multi-year externally-funded projects with budgets between 0.35M and 1M euros and been the primary advisor for 10 completed PhD dissertations.

The WEIGHT-TRAINING Team comprises junior and senior scholars with complementary and overlapping expertise in areas crucial to the project. Table 3 reveals a weakness among the team members currently employed at UiT, namely a lack of expertise in FST theory. This expertise is crucial to authoring the algorithms that manipulate weights in WFST models. This gap will be filled by the ERC-funded PostDoc position, whose main area of expertise will be FST theory, and by hiring a leading expert, Måns Huldén (U Colorado), as a consultant to assist in supervising this activity.

References

•Antonsen, L. 2012. Improving feedback on L2 misspellings – an FST approach. Proceedings of the SLTC 2012 workshop on NLP for CALL, 1–10. •Antonsen, L, S Huhmarniemi, T Trosterud. 2009. Interactive pedagogical programs based on constraint grammar. NEALT Proceedings Series 4. •Antonsen, L, R Johnson, T Trosterud, H Uibo. 2013. Generating modular grammar exercises with finite-state transducers. NEALT Proceedings Series 17: 27–38. •Antonsen, L & LA Janda. forthcoming. Oamastanráhkadusat davvisámi girjjálašvuođas [Possessive constructions in North Saami Literature]. Dieđut. •Bybee, J. 1985. Morphology. Amsterdam: J Benjamins. •Corbett, GG. 2015. Morphosyntactic complexity: a typology of lexical splits. Language 91, 145-193. •Droste, M, W Kuich & H Vogler (eds.) 2009. Handbook of Weighted Automata. Berlin: Springer-Verlag. •Freeman, MMR. 2000. Endangered Peoples of the Arctic: Struggles to Survive and Thrive. Greenwood Press. •Harrison, KD. 2007. When Languages Die. Oxford: Oxford U Press. •Janda, LA. 1995. Unpacking Markedness. In Linguistics in the Redwoods: The expansion of a new paradigm in Linguistics, 207-233 ed. by E Casad. Berlin: Mouton de Gruyter. •Janda, LA. 2006. Cognitive Linguistics. Glossos 8. http://www.seelrc.org/glossos/. •Janda, LA. 2007. Inflectional morphology. In Handbook of Cognitive Linguistics, D Geeraerts & H Cuyckens (eds.), 632-649. Oxford: Oxford U Press. •Janda, LA. 2010. Cognitive Linguistics in the Year 2010. International Journal of Cognitive Linguistics 1, 1-30. •Janda, LA. 2015. Cognitive Linguistics in the Year 2015. Cognitive Semantics 1, 131-154. •Janda, LA & O Lyashevskaya. 2011. Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian. Cognitive Linguistics 22, 719-763. •Janda, LA & V Solovyev. 2009. What Constructional Profiles Reveal About Synonymy: A Case Study of Russian Words for SADNESS and HAPPINESS. Cognitive Linguistics 20, 367-393. •Johnson, R, L Antonsen, T Trosterud. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. NEALT Proceedings Series 16: 59–71. •Joseph, JE & F Newmeyer. 2012. All Languages Are Equally Complex: The rise and fall of a consensus. Historiographia Linguistica 39, 341-368. •Karlsson, F. 1985. Paradigms and word forms. Studia gramatyczne VII. Ossolineum. 135–154. •Lewis, MP, GF Simons, CD Fennig, eds. 2015. Ethnologue: Languages of the World. Dallas: SIL International. online: http://www.ethnologue.com/. •Lupyan, G & R Dale. 2010. Language Structure Is Partly Determined by Social Structure. PLoS ONE 5, e8559. •McCarthy, JJ. 2005. Optimal paradigms. In LJ Downing, TA Hall & R Raffelsiefen (eds.), Paradigms in phonological theory, 170–210. Oxford: Oxford U Press. •McWhorter, J 2011. Linguistic Simplicity and Complexity: Why Do Languages Undress? Berlin: De Gruyter Mouton. •Nesset, T & LA Janda. 2010. Paradigm structure: evidence from Russian suffix shift. Cognitive Linguistics 21, 699-725. •Nickel, KP & P Sammallahti. 2011. Nordsamisk grammatikk. Karasjok: Davvi Girji. •Roche, E. & Y Schabes (eds.) 1997. Finite-State Language Processing. Cambridge, MA: MIT Press. •Schiller, A. 2005. German Compound Analysis with wfsc. Proceedings of the Fifth International Workshop of Finite State Methods in Natural Language Processing. •Trosterud, T. 2004. Homonymy in the Uralic Two-Argument Agreement Paradigms. PhD dissertation, UiT. •Trudgill, P. 2011. Sociolinguistic Typology. Social Determinants of Linguistic Complexity. Oxford: Oxford U Press. •Tyers, FM, L Wiechetek & T Trosterud. 2009. Developing Prototypes for Machine Translation between Two Sami Languages. Proceedings of the 13th Annual Conference of the EAMT, 120-127. •UNESCO Ad Hoc Expert Group on Endangered Languages. 2003. International Expert Meeting On UNESCO Programme Safeguarding of Endangered Languages. Paris. •Unhammer, K & T Trosterud. 2009. Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål. JA Pérez-Ortiz, F Sánchez-Martínez, FM Tyers (eds.) Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, 35–42. •Zipf, GK. 1949. Human behavior and the principle of least effort. Reading, MA: Addison-Wesley.

Electronic references

[1] http://giellatekno.uit.no/index.eng.html
[2] http://sanit.oahpa.no
[3] http://www.arcticlanguages.com/the-project.php#self
[4] http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/news/recommendations_action_plan_atlas_languages.pdf

Section b. Curriculum Vitae

PERSONAL INFORMATION

Janda, **Laura Alexis**
Nationality: USA, Date of birth: November 23, 1957
URL for website: http://ansatte.uit.no/laura.janda/
ORCID: orcid.org/0000-0001-5047-1909
Languages: English, Russian, Czech, Norwegian, North Saami; reading ability in all Slavic languages

EDUCATION

1984	PhD in Slavic Linguistics, Department of Slavic Languages, UCLA, USA
1980	Master in Slavic Linguistics, Department of Slavic Languages, UCLA, USA
1979	AB Cum Laude in Slavic Languages and Literatures & Certificate of Proficiency in Russian Studies, Princeton U

CURRENT POSITION

2008 – present	Professor of Russian Linguistics Faculty of Humanities, Social Sciences and Education at UiT, The Arctic University of Norway. Teaching: Russian & Linguistics, supervision of postdocs, PhD and MA students.

PREVIOUS POSITIONS

1991 – 2007	Professor of Slavic Languages, Department of Slavic Languages, UNC Chapel Hill, USA Teaching: Czech, Russian & Linguistics, supervision of PhD and MA students
1985 – 1991	Assistant Professor of Russian, Department of Foreign Languages, Literatures and Linguistics, University of Rochester, USA Teaching: Russian & Linguistics

MAJOR FELLOWSHIPS AND AWARDS

2007 – 2015	Three grants from Norwegian Research Council
2011	Award for Best Researcher at the Faculty of Humanities, Social Sciences and Education, University of Tromsø for outstanding research and publications.
2011 – 2012	Grant from Centre for Advanced Study at the Norwegian Academy of Science and Letters
2005	American Association of Teachers of Slavic and East European Languages Book Prize
2004, 2005	Grant from National Science Foundation, USA
2003	National Council of Organizations of Less Commonly Taught Languages Award for a distinguished career in Less Commonly Taught Languages
1999 – 2006	Three grants from Title VI Dept of Education, USA
1999 – 2003	Grant from National Security Education Program, USA
1992 – 1994	Grant from Joint Council on Eastern Europe of the American Council of Learned Societies and the Social Science Research Council
1987	Fulbright Research Fellowship

INSTITUTIONAL RESPONSIBILITIES

2013 – 2017	Faculty Council Member, UiT, The Arctic University of Norway
1999 – 2003	Faculty Council Member, UNC Chapel Hill, USA
	PLUS: Committee memberships, too numerous to name.

MAJOR ACHIEVEMENTS

My first book (1986, 105 citations) was on the leading edge in the field of cognitive linguistics and the first to show the radial category semantic structure of Russian prefixes. My monograph on case semantics (1993, 188 citations) demonstrated the same structure for the grammatical category of case, inspiring two research-based handbooks on case for advanced learners of Czech and Russian. In 1996 my book Back from the brink (60 citations) proposed a new pathway for analogical development for morphological markers that rebounded from near extinction to become highly productive. In 2004 I offered innovative insights into the metaphorical structure of Russian aspect, and this spilled over into creative new approaches to teaching. My aspectual clusters model came out in 2007 (86 citations), giving a typology of perfective types in Russian. In 2012 I made the novel proposal that Russian prefixes should be considered a verb classifier system, and this proposal is extended to all types of perfectives and all Slavic languages in a forthcoming article in *Lingua*. In

a series of articles 2009–2014 together with co-authors I launched various types of linguistic profiling: grammatical profiling, constructional profiling, semantic profiling, and radial category profiling.

COMMISSIONS OF TRUST

2008 –present	Associate Editor of *Cognitive Linguistics* (Board Member 1989–2005)
2013	Guest Editor of *Journal of Slavic Linguistics* and *Russian Linguistics*
2007 – 2011	President of the International Cognitive Linguistics Association
2006	Co-founder of Slavic Linguistics Society (Member of Executive Board 2012–present)
2000 – 2008	Co-editor of *Glossos*
2006 – present	Member of Editorial Board of the *Journal of Slavic Linguistics*
2003 – present	Member of Editorial Board of the *Slavic and East European Journal*
2000 – 2003	President and co-founder of the Slavic Cognitive Linguistics Association
2000 – 2002	Vice President of the American Association of Teachers of Slavic and East European Languages
1996 – 1999	Editor of *Czech Language News*
1994 – 1996	President of the North American Association of Teachers of Czech

PLUS: large volume of reviewing for numerous scholarly journals (USA and Europe), academic promotions (USA, Great Britain, France), and grant proposals (Great Britain, USA, Belgium, Poland), and memberships in many scientific societies.

COLLABORATIONS THAT HAVE RESULTED IN JOINT PUBLICATIONS

Lene Antonsen, UiT Norway: *4 articles on North Saami*
Berit Anne Bals Baal, UiT Norway: *2 articles on North Saami*
Harald Baayen, U Tübingen Germany: *2 articles on statistical modeling of Russian rival forms*
Steven Clancy, Harvard U USA: *2 research-based handbooks + 1 article on Czech & Russian case*
Stephen M. Dickey, U Kansas USA: *1 edited anthology + 2 articles on Russian aspectual morphology*
Dagmar Divjak, U Sheffield UK: *3 articles on Russian grammatical constructions*
Hanne Eckhoff, UiT Norway: *3 articles on grammatical profiles of verbs in Old Church Slavonic*
Anna Endresen, UiT Norway: *1 research monograph + 4 articles on Russian aspectual morphology*
Victor Friedman, U Chicago USA: *1 article on Macedonian historical phonology*
Marcin Grygiel, U Rzeszów Poland: *1 edited anthology + 1 article on Slavic cognitive linguistics*
Agata Kochanska, U Warszawa Poland: *1 article on Slavic cognitive linguistics*
John Korba, U Kansas USA: *1 article on Russian aspect*
Julia Kuznetsova, UiT Norway: *1 research monograph + 2 articles on Russian aspectual morphology*
Olga Lyashevskaya, HSE Moscow Russia: *1 research monograph + 6 articles on Russian aspect*
Anastasia Makarova, UiT Norway: *1 research monograph + 4 articles on Russian aspectual morphology*
Tore Nesset, UiT Norway: *1 research monograph, 2 edited anthologies + 11 articles on various topics in Slavic linguistics*
Svetlana Sokolova, UiT Norway: *1 research monograph + 2 articles on Russian aspectual morphology*
Valery Solovyev, U Kazan Russia: *1 article on Russian constructional profiles*
Francis Steen, UCLA USA: *1 article on Russian semantics*
Charles E. Townsend, Princeton U USA: *2 research-based handbooks + 1 article on Russian & Czech*
Mark Turner, Case Western Reserve U USA: *1 article on Russian semantics*
Steven Franks, Ronald Feldstein Indiana U USA: *1 edited anthology*
Petr Sgall, František Čermák, Eva Hajičová, Jiří Hronek, Henry Kučera, Věra Schmiedtová, Jaroslav Suk, Charles U Prague Czech Republic: *1 article on Czech linguistics*

On-Going Grants (External Funding Only)

Project Title	Funding source	Amount (Euros)	Period	Role of the PI	Relation to current ERC proposal
Birds & Beasts: Shaping Events in Old Russian	Norwegian Research Council	1M	2013-2016	co-PI	Focus on the historical development of aspect in Russian and development of corpus resources for Old Russian; very little overlap, if any. Results: empirical approaches to long-standing theoretical controversies.

Prior Grants (Major External Funding Only)

Project Title	Funding source	Amount (Euros)	Period	Role of the PI	Relation to current ERC proposal
Neat theories, messy realities: How to apply absolute definitions to gradient phenomena	Norwegian Research Council	1M	2011-2014	PI	Innovative approach to allomorphy (multiple forms for one meaning); in a sense the converse of the current proposal; no overlap.
Time is Space: Unconscious Models and Conscious Acts	Centre for Advanced Study at the Norwegian Academy of Science and Letters	0.4M	2011-2012	co-PI	Focus on linguistic expression of time; no overlap. Resulted in two anthologies.
Exploring Emptiness: Russian Verbal Morphology and Cognitive Linguistics	Norwegian Research Council	0.7M	2007-2011	co-PI	Focus on prefixes and suffixes traditionally assumed to be semantically empty; some overlap with regard to morphological forms. Resulting monograph and series of articles detail extensive evidence that Russian prefixes always bear meaning.
Matter Matters: A MediaModule for 'The Aspect Book for Russian'	National Science Foundation, USA	85K	2004-2005	PI	Focus on innovative multi-media approach to teaching Russian aspect; no overlap.
Slavic and East European Language Resource Center	Title VI Dept of Education, USA	0.7M	2002-2006	co-PI	Focus on language resources and workshops for instructors; no overlap.
Center for Slavic, Eurasian, and East European Studies	Title VI Dept of Education, USA	0.5M	2000-2003	co-PI	Focus on area studies programming; no overlap.
Slavic and East European Language Resource Center	Title VI Dept of Education, USA	0.44M	1999-2002	co-PI	Focus on language resources and workshops for instructors; no overlap.
Institutional Award to launch new MA in Russian/East European Studies	National Security Education Program, USA	0.35M	1999-2003	PI	Establishment of an MA degree program in area studies; no overlap.
Center for Slavic, Eurasian, and East European Studies	Title VI Dept of Education, USA	0.5M	1997-2000	co-PI	Focus on area studies programming; no overlap.

Research in East European Studies	Joint Council on Eastern Europe of the American Council of Learned Societies and the Social Science Research Council, USA	50K	1992-1994	PI	Focus on historical morphology of Slavic languages; very little overlap. Resulted in a monograph proposing an innovative approach to historical analogy.
Research Fellowship, Charles U., Prague, Czechoslovakia	Fulbright Foundation, USA	30K	1987	PI	Focus on semantics of case in Czech and Russian; very little overlap. Resulted in a research monograph demonstrating that case semantics can be modeled in terms of radial categories.

Applications (External Funding Only)

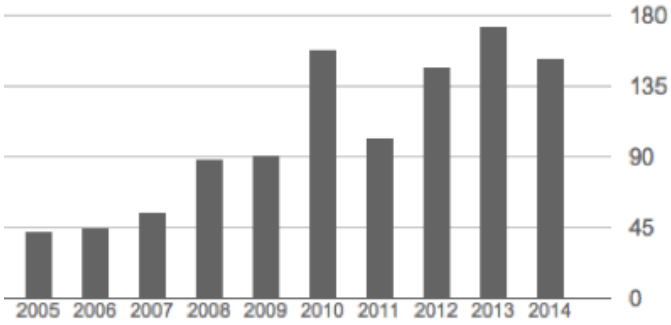
Project Title	Funding source	Amount (Euros)	Period	Role of the PI	Relation to current ERC proposal
WEIGHT-TRAINING	Norwegian Research Council -- Toppforsk	2M	2016-2020	PI	This is essentially the same proposal and will be submitted at the same time. If both the ERC and the Toppforsk proposals are funded, the ERC project will take priority and the Toppforsk proposal will be redesigned to fund a complementary project that does not overlap with the ERC project.
Ambiguity	Norwegian Research Council -- Centres of Excellence Scheme	TBA	2017-2021	PI	This proposal has not yet been written, but will be submitted in November 2015. It will be for a larger project that might overlap with the ERC project. However, if both are funded, the NRC project will be modified so that the projects are complementary.

Section c. Ten-year track-record

1. Articles
Total articles published in 2005–2014: 64 Total articles published in career: 109
Articles scheduled to be published in 2015 or under review: 10
Selected articles republished in Polish, Czech, Russian, and Portuguese.
Citation record in Google Scholar: Total citations = 1467, h-index = 18, i10-index = 37

Ten representative articles published in 2005–2014
(I am first or single author for all 10 articles. Note that ordering of authors is often alphabetical. Numbers of citations are taken from Google Scholar, but this source does not properly represent works I have published in other languages, such as Czech, Russian, and North Saami.)

1. "Aspectual clusters of Russian verbs", *Studies in Language* 31:3 (2007), 607–648. [Citations: 85. This article established an essential classification of types of perfective verbs in Russian that has been adopted by many scholars.]
2. "Cognitive Linguistics". Published in *Glossos* v. 8, 2006 at http://www.seelrc.org/glossos/. 60pp. [Citations: 50. This article has long served as the introductory text in courses in cognitive linguistics. It was updated and republished in both 2010 and 2015.]
3. "What Constructional Profiles Reveal About Synonymy: A Case Study of Russian Words for sadness and happiness", co-authored with Valery Solovyev. *Cognitive Linguistics* 20:2 (2009), 367–393. [Citations: 49. This article debuts the constructional profiling method for probing the relationships among near-synonyms.]
4. "Metonymy in word-formation". *Cognitive Linguistics* 22:2 (2011), 359–392. [Citations: 21. This work demonstrates that metonymy is a motivating factor in word-formation. A critique and follow-up rebuttal were also published.]
5. "Motion Verbs and the Development of Aspect in Russian". *Scando-Slavica* 54 (2008), 179–197. [Citations: 18. This article proposed that motion verbs, usually considered aspectually deviant, are actually prototypical in the Russian aspect system.]
6. "*Xoxotnul, sxitril*: The relationship between semelfactives formed with *-nu-* and *s-* in Russian", co-authored with Stephen M. Dickey. *Russian Linguistics*, 33: 3 (2009), 229–248. [Citations: 17. This article presents the discovery of a relationship between verb classes and the *-nu-* and *s-* prefix which behave like allomorphs.]
7. "Taking Apart Russian RAZ-", co-authored with Tore Nesset. *Slavic and East European Journal* 54:3 (2010), 476–501. [Citations: 17. This article details a new method for semantic analysis of polysemous affixes.]
8. "Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian", co-authored with Olga Lyashevskaya. *Cognitive Linguistics* 22:4 (2011), 719–763. [Citations: 14. This article presents a creative application of statistical measures to address long-standing controversies about Russian aspect.]
9. "Old Church Slavonic *byti* Part One: Grammatical Profiling Analysis and Part Two: Constructional Profiling Analysis", co-authored with Hanne M. Eckhoff and Tore Nesset. 2014. *Slavic and East European Journal* 58.3, 482–525. [Citations: 1. This work presents a creative empirical approach to a long-standing controversy over the status of the 'be' verb in Old Church Slavonic.]
10. "Russkie pristavki kak sistema glagol'nyx klassifikatorov". *Voprosy jazykoznanija* 6 (2012), 3–47. [Citations: 0 in Google Scholar, but this article sparked a lively debate in scholarly circles because it detailed a bold typological claim that Russian prefixes are a verb classifier system. Two critiques and follow-up rebuttals were also published and an extension of this claim is forthcoming in *Lingua*.]



Citations per year in Google Scholar

Total number of books published in career: 18. Breakdown: 4 research monographs (total citations for research monographs: 358), 4 research-based handbooks, 6 edited or co-edited anthologies, 1 anthology of own work, 2 translations by others into German and Korean of a research-based handbook, 1 translation I made of a work of fiction from North Saami into English.

Research monograph in 2005–2014

Why Russian aspectual prefixes aren't empty: prefixes as verb classifiers. 2013. Janda as first author; co-authored with Anna Endresen, Julia Kuznetsova, Olga Lyashevskaya, Anastasia Makarova, Tore Nesset, Svetlana Sokolova. Bloomington, IN: Slavica Publishers. 227pp. [Citations: 8. Based on extensive empirical evidence, Janda and co-authors assert that Russian prefixes, long assumed to be semantically empty when forming verb pairs, in fact always express meaning.]

3. Invited presentations to peer-reviewed, internationally established conferences and/or international advanced schools 2005–2014

Of 164 scholarly presentations delivered in 2005–2014, 78 were invited lectures delivered in the following countries: Norway, Czech Republic, Hungary, Spain, Finland, Poland, USA, Croatia, Sweden, Belgium, Russia, Estonia, China, Denmark, Korea, and Germany. 33 plenary lectures at international conferences; examples include: XLI Kielitieteen päivät / 41st Finnish Conference of Linguistics at the University of Turku, Finland (2014); Jazyk i metod. Russkij jazyk v lingvističeskix issledovanijax XXI veka, at the Jagiellonian University in Kraków, Poland (2014); Quantitative Investigations in Theoretical Linguistics conference in Leuven, Belgium (2013); Slavic Linguistic Society conference, Zadar, Croatia (2009); SKY, Finnish Linguistics Society, Helsinki (2009); International Cognitive Linguistics Conference in Kraków, Poland (2007); International Conference on the Russian National Corpus, Moscow, Russia (2007); Conference on Interdisciplinarity in Cognitive Science Research, Russian State University for the Humanities, Russia, (2012); Festival of Slavic Languages at Irkutsk State University, Russia (2012); Third Finnish-Estonian Cognitive Linguistics Conference, Estonia (2011); 10 lectures presented at 7 universities in Beijing, China in Eminent Linguists Lecture Series of the China International Forum on Cognitive Linguistics (2011).

4. Organization of international conferences 2005–2014

The largest international conferences I have participated in organizing were the International Cognitive Linguistics Conferences in Xian, China (2011) and Edmonton, Canada (2013), both of which had several hundred attendees. I am involved in organizing an average of 1-2 conferences per year.

5. Prizes & Awards 2005–2014

Best Researcher at the Faculty of Humanities, Social Sciences and Education, University of Tromsø (2011); Book Prize from the American Association of Teachers of Slavic and East European Languages (2005). Major grants from: Norwegian Research Council (3X: 2007-2015), Centre for Advanced Study at the Norwegian Academy of Science and Letters (2011-2012), National Science Foundation of USA (2005), Title VI Dept of Education of USA (2002-2006).

6. Major contributions to the early careers of excellent researchers

I have been the primary advisor for 3 post-docs (2008-present),10 PhD dissertations (5 completed in 2005–2014) and 19 MA theses (7 completed in 2005–2014), and have mentored 9 foreign scholars. Most notable among my mentees are: Dagmar Divjak (Reader & Head of Dept, U of Sheffield, British Academy Fellow), Olga Lyashevskaya (Professor, Higher School of Economics, Moscow), Steven Clancy (Senior Lecturer, Harvard U), Anne Stepan (Senior Analyst [ranks between Colonel and General], US Dept of Defense), Patrick Murphy (Assoc. Professor, U Maryland and Linguist for US Federal Government), Hyug Ahn (Asst. Professor, Sung-Kyun-Kwan U, Seoul, Korea), Svetlana Sokolova (Assoc. Professor, UiT), Anna Endresen (Lecturer, UiT), and Wojciech Lewandowski (Marie-Curie Postdoctoral Fellow, U Copenhagen).