Morphological Aspects of Computer-Driven Elicitation of Knowledge about Any Language

Sergei Nirenburg and Marjorie McShane
University of Maryland Baltimore County

How to create NLP resources for a new language?

By hand

By parameterizing:

- Translation-based methods (e.g., Probst, Levin et al.)
  - Easy for informants
  - Coverage a problem
- ML (quick ramp-up competitions; parallel corpora; learning morphology)
  - Quality a problem
  - Rules are not human interpretable
- Interactive knowledge elicitation system
  - Lots of choices which determine ultimate form, coverage, etc., of knowledge
  - This talk focuses on some of the choices and their implications in the realm of morphology

Boas Knowledge elicitation system

- Supports computational field linguistics
- Extracts knowledge about any language from a non-expert informant
- No knowledge engineer involved
- English is assumed as the language of HCI
- Mixed-initiative knowledge elicitation strategy
- System is supplied with cross-linguistically motivated inventory of parameters and values
- The “signature” of a language is how it realizes linguistic parameters

Examples of parameters and values

Case: nominative, accusative, dative…

Number: singular, plural, dual, paucal…

Tense: past, present, future…

Aspect: progressive, simple…

Grammatical role: subject, direct object…

Agreement: subject-verb, noun-adjective…

Open-class lexical meanings: word, phrase, morpheme…

Closed-class lexical meanings: word, phrase, morpheme, feature…

…

Some examples of phenomena we must treat

From:

McShane, Marjorie and Sergei Nirenburg. 2004. Parameterizing the Space of Lexical and Grammatical Meaning Across Languages. *Machine Translation*. 18(2) : 129-165.

French

Étudie-_	**t-_**	**elle**	**maintenant? –**	**Non,**	**elle**	**m'_**	**attend**	**à**
study$_{3.SG.PRES}$	PARTICLE	she$_{3.SG.NOM}$	now	no	she$_{3.SG.NOM}$	me$_{1.SG.OBJ.}$	waits$_{3.SG.PRES}$	at

l'_ **université.**

the university$_{MASC.SG.}$

'Is she studying now?' 'No, she's waiting for me at the university.'

German

Nach	**Angaben**	**der**	**brit_isch_en**	**Regierung**
according-to	statement$_{FEM.PL.DAT}$	of$_{FEM.SG.GEN}$	British$_{FEM.SG.GEN}$	administration$_{FEM.SG.GEN}$

schlug	**Blair**	**in**	**einem**	**Brief**	**an**	**die**
hit	Blair	in	a$_{MASC.SG.DAT}$	letter$_{MASC.SG.DAT}$	to	the$_{MASC.PL.ACC}$

Regierungs_	**chefs**	**der**	**Nato-_**	**Staaten**	**und**
governments	heads$_{MASC.PL.ACC}$	of$_{MASC.PL.GEN}$	NATO	states$_{MASC.PL.GEN}$	and

an	**den**	**russ_isch_en**	**Präsidenten**	**Wladimir**	**Putin**
to	the$_{MASC.SG.ACC}$	Russian$_{MASC.SG.ACC}$	President$_{MASC.SG.ACC}$	Vladimir	Putin

die	**Bildung**	**eines**	**neuen**
the$_{FEM.SG.ACC}$	formation$_{FEM.SG.ACC}$	a$_{MASC.SG.GEN}$	new$_{MASC.SG.GEN}$

Russland-_	**Nord_**	**atlantik_rats**		**vor.**
Russia	North	Atlantic	council$_{MASC.SG.GEN}$	in-front-of

'According to statements by the British administration, Blair, in a letter to the heads of governments of the NATO states and to Russian president Vladimir Putin, suggested the formation of a new Russia-North-Atlantic Council.'

Russian

Ja	**by**	**udarila**	**ego**	**palkoj.**
$I_{1.SG.NOM}$	CONDITIONAL	$hit_{3.SG.FEM.PAST}$	$him_{ACC.SG.MASC.}$	$stick_{INSTR.SG.FEM}$

'I would have hit him with a stick.'

Ukrainian

a.	**Ja**	**budu**	**govoryty**	**tyxše,**	**niž**	**ty.**
	$I_{NOM.SG}$	will	$speak_{INFIN}$	quieter	than	$you_{NOM.SG}$
b.	**Ja**	**govorytymu**	**tyxše,**	**niž**	**ty.**	
	$I_{NOM.SG}$	$speak_{1.SG.FUT}$	quieter	than	$you_{NOM.SG}$	

'I will speak more softly than you.'

Polish

a.	**My_**	**śmy**	**znowu**		**wczoraj**		**poszli**		**do**	**parku.**
	we $_{1.NOM.PL}$	**1.PL**	again		yesterday		went$_{3.PL}$		to	park$_{GEN.SG}$
b.	**My**		**znowu_**	**śmy**	**wczoraj**		**poszli**		**do**	**parku**
	we $_{1.NOM.PL}$		again	**1.PL**	yesterday		went$_{3.PL}$		to	park$_{GEN.SG}$
c.	**My**		**znowu**		**wczoraj_**	**śmy**	**poszli**		**do**	**parku.**
	we $_{1.NOM.PL}$		again		yesterday	**1.PL**	went$_{3.PL}$		to	park$_{GEN.SG}$
d.	**My**		**znowu**		**wczoraj**		**poszli_**	**śmy**	**do**	**parku.**
	we $_{1.NOM.PL}$		again		yesterday		went$_{3.PL.}$	**1.PL**	to	park$_{GEN.SG}$

'We went to the park again yesterday.'

Turkish

(ben)	**Hasan_a**		**bavul_**	**u**	**taşı_**	**t_**	**ti_**	**m**
I	Hasan	DAT	suitcase	ACC.SG.	carry	CAUS.	PAST	1.SG

'I made Hasan carry the suitcase'

Persian

Sarma_	**ye**	**shadid**	**Ali**	**ra**	**kosht.**
$cold_{SG.}$	EZAFE	severe	Ali	POSTPOSITION .OBJ.MARKER	$kill_{PAST}$

'A severe cold killed Ali.'

Hebrew

keshe_	**pagash_**	**ti_**	**h_**	**a**
when	met	I	you	MASC.

'when I met you'

a. Irish

sráid ~ **an tsráid**
street ~ the street

b. Bulgarian

more_ **to**
$\text{sea}_{\text{NEUT.SG}}$ $\text{the}_{\text{NEUT.SG}}$ 'the sea'

c. Czech

ne_ **znáte**
not $\text{know}_{\text{2.PL.PRES}}$ '(you) don't know'

d. Tagalog

bulaklak ~ **magbu_bulaklak**
flower ~ flower vendor

A text element may contain **one stem** (Fr. *elle*; Tur. *Hasana*) or **multiple stems** (Fr. *m'attend*; G. *Russland-Nordatlantikrats*).[11]

Stems may represent:

- **open-class elements**—nouns (G. *Angaben*; Ir. *sráid*), verbs (Tur. *taşıttim*; Per. *kosht*), adjectives (G. *neuen*; U. *tyxše*), adverbs (Fr. *maintenant*);
- **closed-class elements**—pronouns (Fr. *elle*; Pol. *my*), conjunctions (G. *und*; U. *niž*), prepositions (Fr. *à*; Ger. *der, in, an, vor*; Pol. *do*), articles (Fr. *l'*; Ger. *den, die*), etc.;
- **inflectional elements**—auxiliaries (U. *budu*, R. *by*), postpositions (Per. *ra*).
- **onomasticon elements**—proper nouns (Ger. *Wladimir Putin*), proper adjectives (Ger. *britischen*).

Open-class stems may be inflected using **synthetic** flective inflection (Fr. *Étudie*; R. *udarila*; G. *Angaben*), **analytical** inflection (U. *budu govoryty*) or **agglutinating** inflection (H. *keshepagashtiha*).

Closed-class stems may also be inflected, often in **suppletive** paradigms (R. *ego*).

Inflection may represent **syntactic** information (Pol. *My* is in the nominative case, indicating that it is a subject) or **lexical** information (R. *palkoj* is instrumental singular, with the instrumental case reflecting the closed-class meaning 'with').

If an element contains multiple stems, the stems may be separated by a **hyphen** (Fr. *Étudie-t-elle*; G. *Nato-Staaten*), an **apostrophe** (Fr. *m'attend, l'université*), or **nothing at all** (G. *Nordatlantikrats, Regierungschefs*).

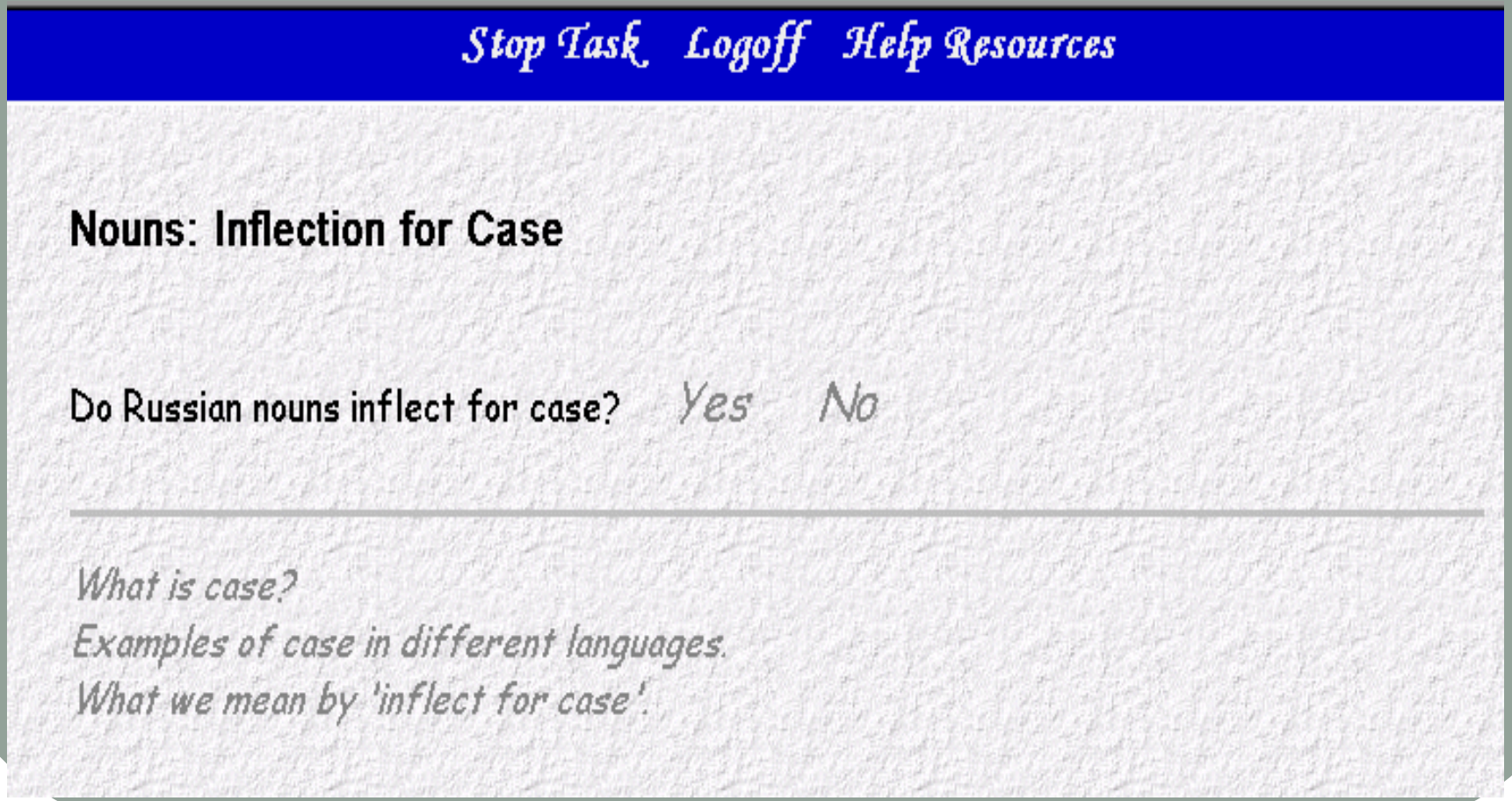
Multi-stem text elements may contain: **two or more open-class stems** (G. *Nato-Staaten*) or a combination of **open-class and closed-class stems** (Fr. *Étudie-t-elle*; H. *keshepagashtiha*).

Derivational word-formation processes that can affect a stem include **compounding** (G. *Nato-Staaten*), **affixal derivation** (Cz. *neznáte*; G. *britischen, russischen*), **reduplication**, or some **combination** of the above (Tag. *magbubulaklak*).

Syntax-level word-formation processes, which are sometimes induced by phonetic reasons, include **insertion of phonetic elements** (Fr. *t* in *Étudie-t-elle*), **affixal realizations of closed-class** items (Fr. *m'attend*; H. *keshepagashtiha*), words formed by inflectional **affix hopping** (Pol. *myśmy, znowuśmy, wczorajśmy*), and syntactically determined **spelling variants** (Ir. *an tsráid*).

How to get an inventory of parameters and values?

Boas uses a series of KE screens like this one:



Choices related to morphology

- Informant: novice or expert?
- For inflectional morphology: paradigms or not?
- How to define a paradigm for purposes of ML?
- How might an informant help to learn inflectional morphology?
- What should a paradigm look like?
- How can the system help to create good paradigms?
- Are analytical forms part of the paradigm?
- Does is matter how many paradigms there are?
- How to elicit irregular forms of open-class items?
- How to elicit agglutinating morphology?

The informant: a linguistic novice or An expert?

If the informant is an expert the system must:

- Initiate the expert into a KE process that is more constrained than that typically used in field linguistics; e.g., must use the expressive means provided in the system and not, say, free prose
- Impose a division of linguistic reality into modules supported by processing engines for which the knowledge is being acquired
- Emphasize that typical cases are most important; not focus all energies on exceptions (no "teratology")
- Coax the expert into carrying out more manual work than he/she might want to do

- If the informant is a novice the system must:
  - Provide extensive pedagogical materials (if we assume no human linguistic guide accompanying the system)
  - Assist the informant in creating generalizations and abstractions
  - Provide redo and refinement capabilities, to the extent possible, with as little work lost as possible (this can get very complex fast: e.g. a person has created inflectional paradigms and realizes he has forgotten a parameter)
  - Help the informant to find a balance between listing (conceptually easier but more time-consuming) and creating abstractions and

For Inflectional morphology, Paradigms or not?

The existence of morphological analyzers for flective languages is practically taken as a given in current NLP systems but it is not, in principle, a necessity.

Listing all inflectional forms in the lexicon might be a better (or additional) option if:

- Labor is cheap
- A language has a lot of irregular forms (e.g., Irish)
- The knowledge engineers have difficulty making linguistic abstractions

Another option would be **to elicit inventories of affixes and morphotactic rules**; however, it would be difficult to develop expressive means that would permit an informant to convey, in a way the program could unambiguously understand, the range of possible

Pros and cons of inflectional paradigms

Pros:

- Can learn morphological rules in well-understood ways
  - Limits listing of forms, saves time
  - Covers new inputs (e.g., newly coined words)

Cons of paradigms

- How to define a paradigm? How many paradigms?
  - Russian grammars typically state that there are four major nominal declensions but Wade (1992) posits over thirty subclasses and Zalizniak (1967), over 70
  - Polish grammars tend to avoid the paradigm issue completely for nouns, giving stem-specific endings for each combination of case, number, gender and virility (Bielec 1998; Kaleta 1998)

How to define paradigm for ML?

Will inherent features, like gender or animacy, be available to the learner/analyzer/generator, e.g., through a lexicon?

Will the ML algorithm permit paradigm bunching based on the stem form?

- E.g., for the **invented** example below, would the learner learn that stems in –dyt have a different NomPl ending than stems in –myt?

padyt$_{\text{Nom.Sg.}}$	padyt**is**$_{\text{Nom.Pl.}}$
romyt$_{\text{Nom.Sg.}}$	romyt**os**$_{\text{Nom.Pl.}}$

Our most sophisticated ML engine

Developed by Kemal Oflazer (Oflazer, Nirenburg, McShane 2001)

Sample inflectional paradigms were compiled into a finite state transducer lexicon and combined with a sequence of morphographemic rewrite rules induced using transformation-based learning

The engine generated as well as analyzed

Permitted learning loop elicitation methodology:

- Informant provides full inflectional forms for one example of each paradigm
- System learns rules
- Informant provides more citation forms as examples
- System generates what it believes to be correct forms
- Informant corrects mistakes; system relearns

Case	Number	
	Singular	Plural
Nominative	telefon	telefony
Vocative	telefon**ie**	telefony
Accusative	telefon	telefony
Genitive	telefon**u**	telefon**ów**
Dative	telefon**owi**	telefon**om**
Locative	telefon**ie**	telefon**ach**
Instrumental	telefon**em**	telefon**ami**

Polish was used as a test case for development of the ML system.

To the left is the primary example of a "bunched" paradigm that ultimately covered 18 word-final consonants and consonant clusters.

Additional provided inflectional forms permitted the learning of many stem-specific variations, including such things as:

- b,p,f,w,m,n,s,z,t,d,st,zm take Loc/Voc Sing ending –ie
- g,k,ch take Loc/Voc Sing ending –u
- Before the Loc/Voc Sing affix, many consonant alternations take place, including $st \rightarrow ść$, $zm \rightarrow źm$, $ł \rightarrow l$, $r \rightarrow rz$, $sł \rightarrow śl$; and others
- Etc!

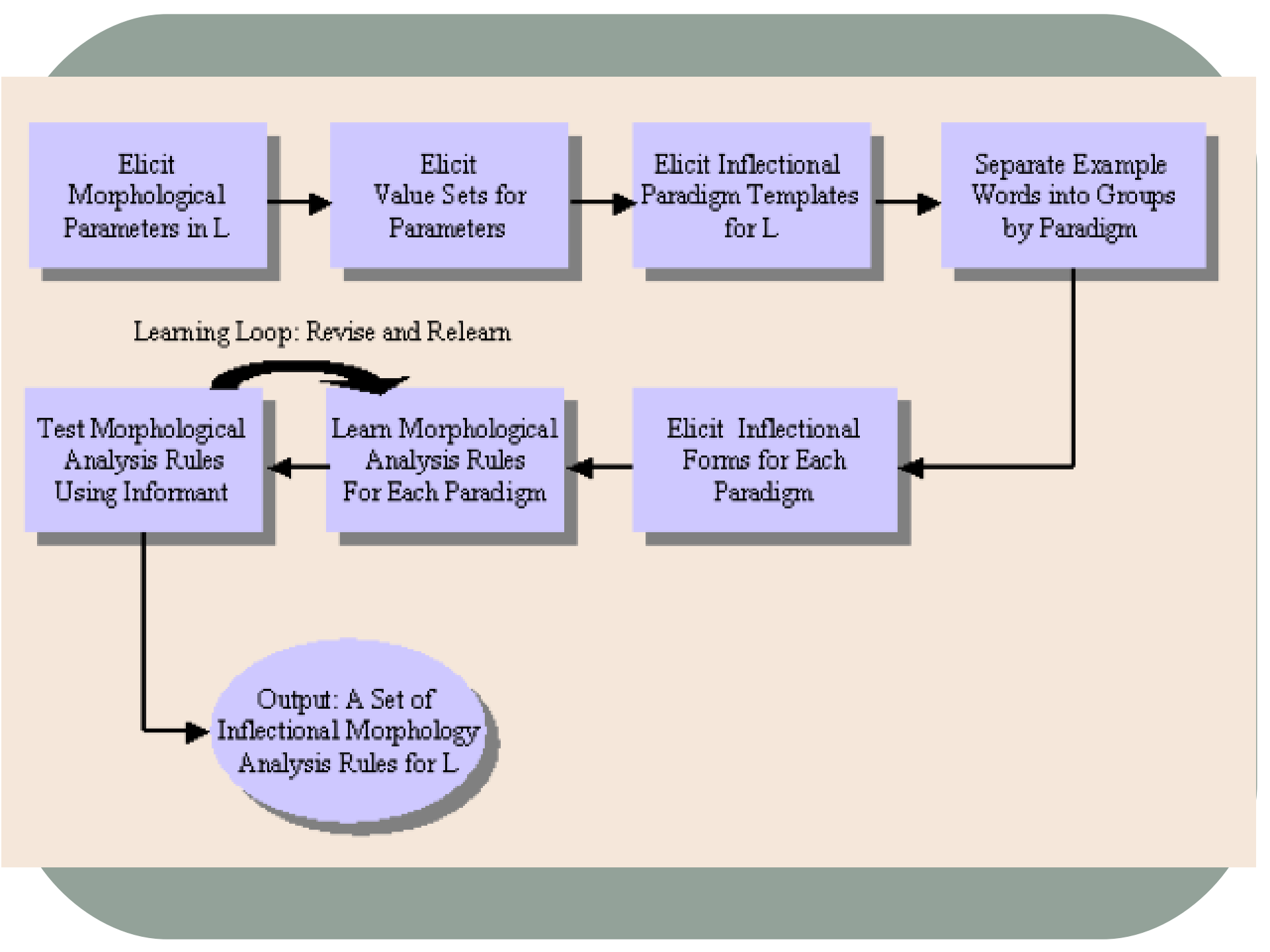
What if the ML algorithm must be simpler?

The approach just described relied on a toolset that ceased to be available at some point

It also involved a learning-loop methodology that involved generation of forms as well as analysis

As an alternative, one can create a simpler ML system that shifts some more work to the informant; e.g.,

- Perhaps all words in every paradigm must inflect exactly the same way
- Perhaps the informant will be asked to explicitly specify the stem for each paradigm
- Perhaps the informant will be required to indicate stem-affix boundaries (segmentation)


Elicit Morphological Parameters in L
Elicit Value Sets for Parameters
Elicit Inflectional Paradigm Templates for L
Separate Example Words into Groups by Paradigm
Learning Loop: Revise and Relearn
Test Morphological Analysis Rules Using Informant
Learn Morphological Analysis Rules For Each Paradigm
Elicit Inflectional Forms for Each Paradigm
Output: A Set of Inflectional Morphology Analysis Rules for L

How might an informant help to generate good inflectional rules?

- Rules could be shown to the informant, if they are or could be made transparent: e.g., a bad rule for dog > dogs would be "lop off the g and add gs"
- The bootstrapping learning loop methodology could be used, as described earlier (generate forms for the user to check)
- The informant could be asked for additional information to guide learning: an inventory of affixes, the stem for each paradigm, known stem alternations, whether or not the language has fleeting vowels, whether the language uses infixation, etc.

In one version of the Boas system, a simplified ML engine was used and informants were asked to indicate the stem.

Selecting the STEM for the *самолет* Paradigm

In the text field below, please indicate the stem of the word самолет - i.e. the core form from which inflectional forms are built (this is *not* necessarily the same as the citation form!). The idea is to help the system to create linguistically valid rules of inflection that can later be generalized to other words of this type.

Stem:	самолет
Citation Form	самолет
Singular Nominative	самолет
Singular Accusative	самолет
Singular Genitive	самолета
Singular Dative	самолету
Singular Instrumental	самолетом
Singular Locative	самолете
Plural Nominative	самолеты
Plural Accusative	самолеты
Plural Genitive	самолетов
Plural Dative	самолетам
Plural Instrumental	самолетами
Plural Locative	самолетах

What should a paradigm look like?

- The more forms a paradigm has, the more important it is to organize them in a way that is convenient for a given informant; the traditional large tables of reference grammars are not necessarily the most user-friendly
- The two following screen shots show a novel layout for a French paradigm
- Emphasis on visualization, organization, putting like forms together for ergonomic reasons and, possibly, to support more efficient ML
- Recent advances in GUI technologies could revolutionize how we think about grammar description.

arriv-

Présent (Ind.)	*Présent (Sub.)*
arrive	
arrives	
arrive	
arrivons	arriv i ons
arrivez	arriv i ez
arrivent	

Arriv-

Imparfait (Sub.)
arrivasse
arrivasses
arrivât
arrivassions
arrivassiez
arrivassent

arriv-

Passé simple (Ind.)	*Imparfait (Ind.)*
arrivai	arrivais
arrivas	arrivais
arriva	arrivait
arrivâmes	arriv i ons
arrivâtes	arriv i ez
arrivèrent	arrivaient

arriver-

Fut. Simple (Ind.)	*Présent (Cond.)*
Arriverai	arriverais
Arriveras	arriverais
Arrivera	arriverait
Arriverons	arriver i ons
Arriverez	arriver i ez
Arriveront	arriveraient

- All single-word forms together (see next page for compound forms)
- Upper left: mostly merge present indicative and subjunctive; visually underscore difference (addition of "I" using spacing or another means)
- Shaded cells in left and right bottom have same endings.

Passé (Sub.)		*Passé composé (Ind.)*		*Plus-que-parfait (Ind.)*	
sois	soyons	suis	sommes	étais	étions
sois	soyez	es	êtes	étais	étiez
soit	soient	est	sont	était	étaient

Passé (Con.) 1		arrivé(e)(s)		*Futur antérieur (Ind.)*	
serais	serions			serai	serons
serais	seriez			seras	serez
serait	seraient			sera	seront

Passé (Con.) 2		*Plus-que-parfait (Sub.)*		*Passé antérieur (Ind.)*	
fusse	fussions	fusse	fussions	fus	fûmes
fusses	fussiez	fusses	fussiez	fus	fûtes
fût	fussent	fût	fussent	fut	furent

- All compound forms have same participial forms of main verb, represented as a hub.
- Shaded boxes show that same forms can represent two meanings.
- All "ser-" forms in one row; all "fus-" forms in another row.

The paradigm layouts in boas

Two main options (these days GUI support would allow much more)

- Simple table with parameters in any order
- Hierarchical layout, with parameters in any order and any number of levels of hierarchy

Simple Table with Number preceding Case.

Singular	Nominative	word
Singular	Genitive	word
Singular	Dative	word
Plural	Nominative	word
Plural	Genitive	word
Plural	Dative	word

Hierarchy with Number preceding Case.

Singular

Nominative	word
Genitive	word
Dative	word

Plural

Nominative	word
Genitive	word
Dative	word

Once the template is created, How can the system help to create good paradigms?

Boas had "fast lane" and "scenic route" for creating paradigms

Fast lane: informant decides how many/which paradigms to create

Scenic route:

- Informant translates a pre-compiled word list and indicates any grammatically relevant inherent features
- The system posits paradigms based on inherent features and word forms (ends with consonant, ends with vowel, etc.)
- Informant decides whether the words in the posited groups inflect similarly; splits or bunches groups accordingly

Scenic route, step 1: Translation and inherent features

Scenic route, Steps 2&3: paradigm guessing, modifying

Are analytical inflectional forms part of the paradigm?

It is conceptually simplest to include these in the paradigm, especially since in some languages synthetic and analytical forms are alternatives: Ukrainian budu pysaty vs. pysatymu

But this complicates the ML algorithm.

The boas solution

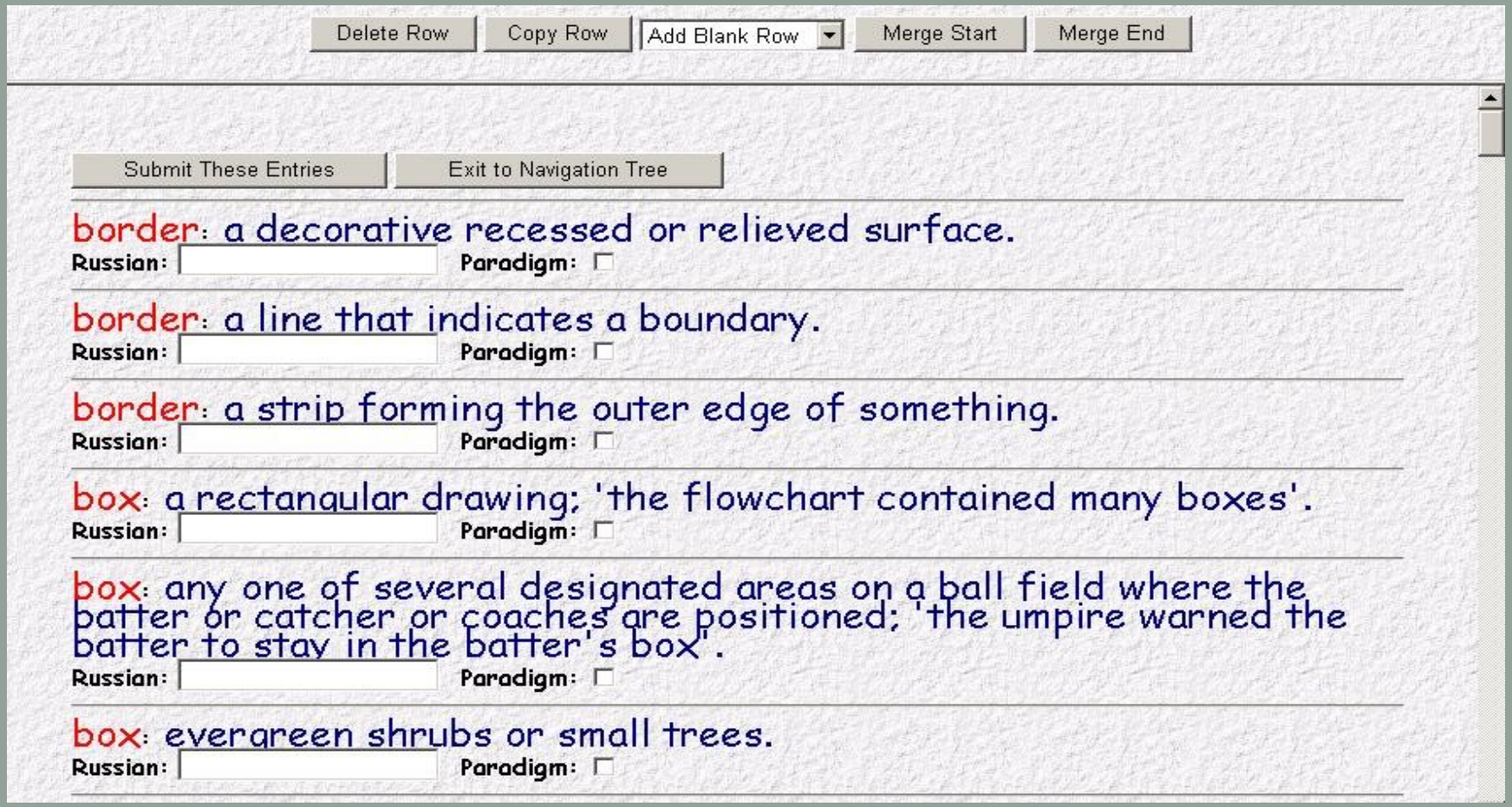
- Once the paradigm template is established, the informant divides table cells up into those that have single-word, multi-word, or both single- and multi-word realizations
- The single-word entities are extracted and dealt with in paradigms in the way described above
- Multi-word entities are sent to a different elicitation module where the informant provides an inventory of auxiliaries and their inflectional forms, if applicable, then links them to the correct forms of the main verb

Does it matter how many paradigms there are?

It does if the informant is expected to manipulate them later on: e.g., if he is expected to assign every open-class word to a paradigm explicitly

If explicit assignment is necessary, then it should at least be made semi-automatic

How to elicit Irregular forms of open class items?



Clicking on "paradigm" takes user to a template to fill out.

How to elicit Agglutinating Inflectional morphology?

Combined agglutinating affix elicitation

Grammatical Number and Person Combined (non-paradigmatically)

Even languages that are highly agglutinating or isolating (i.e., use a separate affix or word for each bit of meaning) tend to combine grammatical *Person* and *Number* in a single affix/word, particularly on verbs. For example, there might be one affix/word for First Person Singular, another for First Person Plural, etc. If Russian combines Person and Number in its inflectional affixes/words, list the relevant affixes/words in the table below, one to a line. (You will have the chance to list separate affixes/words for each of these parameters in later pages.)

For affixes, type a period at the place of attachment:

.suffix
prefix.
circumfix-beginning.circumfix-end
.infix.

Click here for a description and example of each value of Number.
Click here for a description and example of each value of Person.

First Person Singular	
First Person Plural	
First Person Dual	

Where else can inflectional meanings occur?
morphology/syntax interface

Grammatical Functions: Direct Objects

The *direct object* denotes the person or thing that is directly affected by the verbal action (precisely defining direct object can be rather sticky in some languages). It is likely that at least some of the blue noun phrases below are direct objects in Russian, as they are in English. Consider how you would express them in Russian. (Variants are listed using a slash; they are provided only in case different words behave differently for some reason in Russian.)

She likes French novels / onion soup / sunny days / her mother-in-law / him.
She doesn't like French novels / onion soup / sunny days / her mother-in-law / him

She gave a kitten / a doll house / a bicycle / it to her daughter as a birthday gift.
He delivered a letter / it to my cousin.

She would eat steak / ice cream / it every day if she could.
She wouldn't eat steak / ice cream / it if you paid her to.

For those blue noun phrases that are direct objects, how does Russian show that they are direct objects? Please select as many of the following options as are applicable. (Select "word order" only if the direct object **must** occupy a given position with respect to the verb, like direct objects do in English.)

By case-marking.	☐
By the use of a particle, preposition or postposition.	☐
By word order.	☐

More morphological realizations: Closed class lexicon

Means of Realizing Closed-Class Items

Closed-class items may be realized in any of four ways: as a word, a phrase, an affix, or a feature. It is possible that more than one type of realization can be used to translate a given closed-class meaning in your language.

Below are examples of each type of realization.

- **Word Realization**
  The English preposition 'for' is translated by the French word 'pour'.
- **Phrasal Realization**
  The English preposition 'below' is translated by the French phrase 'au-dessous de'.
- **Affixal Realization**
  The English preposition 'the' is translated by the Bulgarian suffixes '-to', '-ta', etc.: more ~ more**to** "sea ~ the sea", staja ~ staja**ta** "room ~ the room".
- **Feature Realization**
  The English preposition 'with' - as in '(hit someone) with a stick' - is translated into Russian by putting the noun in question ('stick') in the instrumental case.

Closed class Lexicon acquisition interface

And still more meanings realized via affixes

What can be exploited in building NLP resources?

Cross-linguistic generalizations: parameters and values, lists of various types, etc.

Technology

- Complex control structures (e.g., permitting redo of partial results: "I forgot one parameter in the 40 paradigms I have already developed!")
- Importing available resources for L (e.g., lexicon)
- Modifying available resources for some language like L (e.g., build Catalan system from Spanish one)