Putting activity models in the driver's seat

Towards a demand-driven NLP architecture for ICALL

Luiz Amaral
University of Victoria

Detmar Meurers
The Ohio State University

EUROCALL 2007
University of Ulster, Coleraine Campus
September 6, 2007

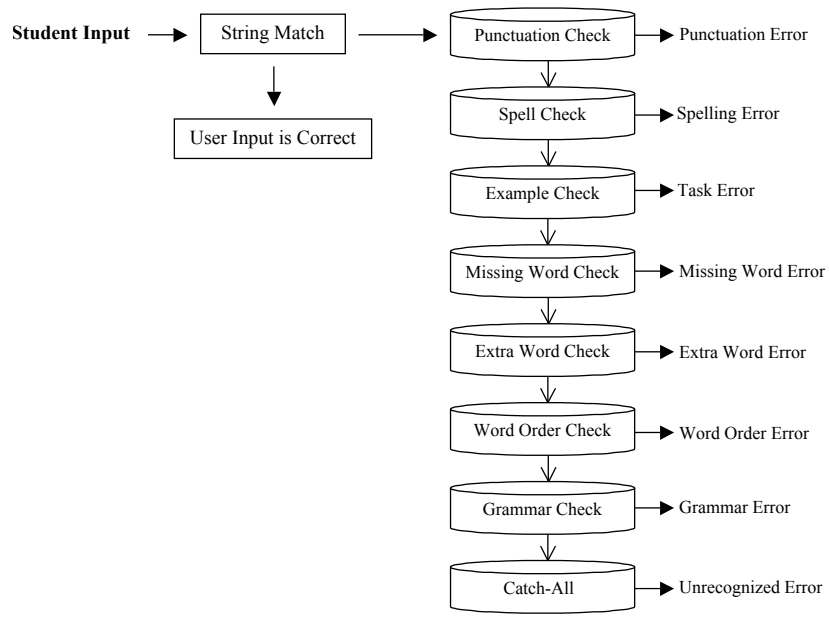
Computer-Assisted Language Learning and Natural Language Processing

- General goal: CALL systems that are able to
  - diagnose errors and provide individualized feedback
  - individually adjust the sequencing of instruction
- Options for achieving this:
  - preenvisage all possible learner inputs and the intended system responses, or
  - automatically analyze the learner input using algorithms and resources from Natural Language Processing (NLP).
- ICALL pursues the latter approach, which
  - requires less hand-coding for each exercise, and
  - in principle supports a wider range of activities
    - including those where preenvisaging all answers requiring feedback is impossible or practically infeasible.

How is NLP integrated into ICALL systems?

- ICALL systems typically process learner input by
  - calling NLP modules in a pre-defined order,
  - transforming one data structure into another and
  - terminating when specific conditions are met.

An Example: German Tutor (Heift and Nicholson 2001)



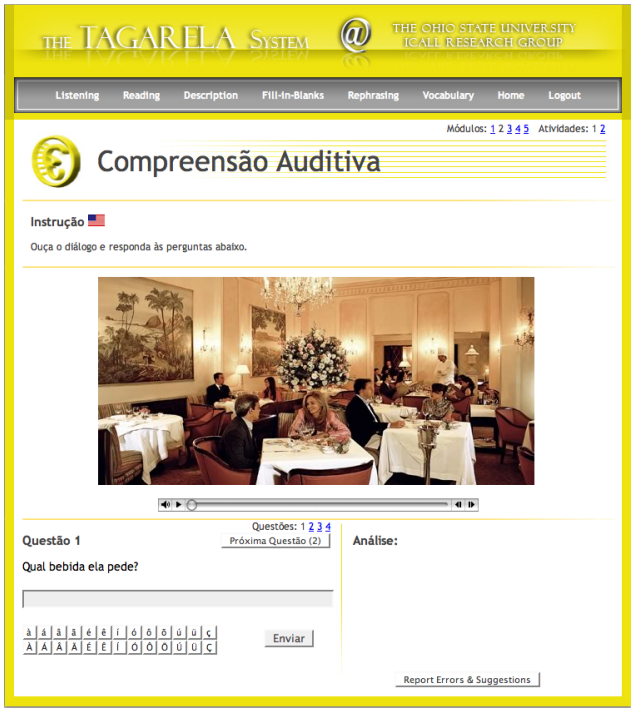
Pipeline architectures in ICALL

Where it's good and where it's suboptimal

- A pipeline architecture works well as long as the system deals with learner input from activity types that are uniform with respect to the required NLP processing.
- What is needed to develop ICALL systems which
  - integrate a wider range of activity types,
  - resulting in learner input of a heterogeneous nature,
  - which should be evaluated on different criteria
  - with the feedback also depending on the learner history?
- ICALL architectures would benefit from
  - viewing the NLP modules as enriching the input with annotations (as in current corpus annotation work), and
  - giving the activity model control over which annotations can or must be provided in a demand-driven architecture

Research Context: The TAGARELA system

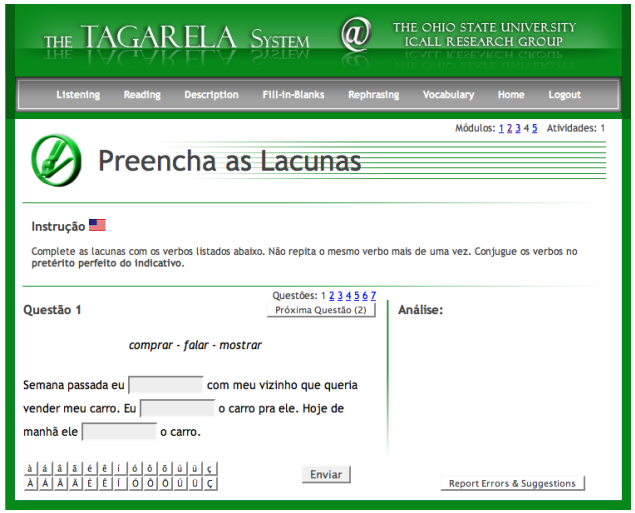
- We are developing TAGARELA, an intelligent workbook accompanying the teaching in the OSU Portuguese Language Program (Amaral and Meurers 2005, 2006)
- TAGARELA offers on-the-spot individualized feedback on spelling, morphological, syntactic and semantic errors.
- It addresses real-life needs identified in interviews with OSU foreign language instructors (Amaral 2004).
  - It provides opportunities for students to practice their listening, reading, and writing skills.
  - Focus: providing on-the-spot feedback for activity types considered useful but typically assigned as homework.



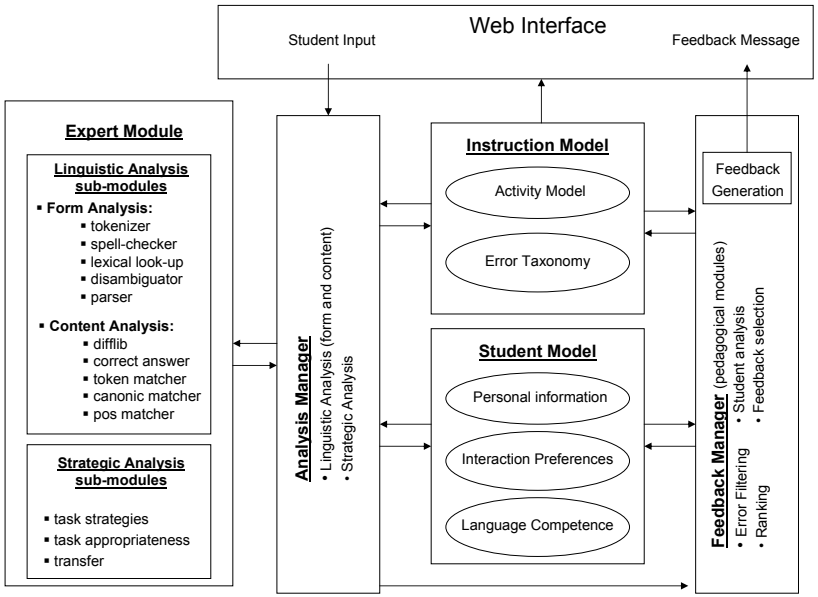

THE TAGARELA System
@
THE OHIO STATE UNIVERSITY
ICALL RESEARCH GROUP
Listening Reading Description Fill-in-Blanks Rephrasing Vocabulary Home Logout
Módulos: 1 2 3 4 5 Atividades: 1
Descrição
Instrução
Descreva a foto usando as palavras apresentadas no exercício e uma das preposições abaixo.
em cima de - entre - embaixo de - ao lado de
Questões: 1 2 3 4
Questão 1
Próxima Questão (2)
Análise:
vaso - mesa
Enviar
Report Errors & Suggestions


THE TAGARELA System
@
THE OHIO STATE UNIVERSITY
ICALL RESEARCH GROUP
Listening Reading Description Fill-in-Blanks Rephrasing Vocabulary Home Logout
Módulos: 1 2 3 4 5 Atividades: 1
Reescreva
Instrução
Escreva uma frase comparando os dois elementos apresentados na tabela. Siga o exemplo abaixo.
ensolarada
cozinha 1: sol todo o dia
cozinha 2: sol só pela manhã
Resposta: A cozinha 1 é mais ensolarada que a cozinha 2.
Questões: 1 2 3 4 5 6 7 8
Questão 1
Próxima Questão (2)
Análise:
caro
apartamento 1: R$150.000
apartamento 2: R$230.000
Enviar
Report Errors & Suggestions


THE TAGARELA System
@
THE OHIO STATE UNIVERSITY
ICALL RESEARCH GROUP
Listening Reading Description Fill-in-Blanks Rephrasing Vocabulary Home Logout
Módulos: 1 2 3 4 5 Atividades: 1
Vocabulário
Instrução
Observe a figura e complete a descrição com as palavras que estão faltando.
Questões: 1 2 3 4 5 6 7 8 9
Questão 1
Próxima Questão (2)
Análise:
No banheiro tem .
Enviar
Report Errors & Suggestions


THE TAGARELA System
@
THE OHIO STATE UNIVERSITY
ICALL RESEARCH GROUP
Listening Reading Description Fill-in-Blanks Rephrasing Vocabulary Home Logout
Módulos: 1 2 3 4 5 Atividades: 1
Preencha as Lacunas
Instrução
Complete as lacunas com os verbos listados abaixo. Não repita o mesmo verbo mais de uma vez. Conjugue os verbos no pretérito perfeito do indicativo.
Questões: 1 2 3 4 5 6 7
Questão 1
Próxima Questão (2)
Análise:
comprar - falar - mostrar
Semana passada eu com meu vizinho que queria vender meu carro. Eu o carro pra ele. Hoje de manhã ele o carro.
Enviar
Report Errors & Suggestions

Overall system architecture for TAGARELA



NLP analysis modules in TAGARELA

- Form Analysis:
  - tokenizer: takes into account specifics of Portuguese (cliticization, contractions, abbreviations)
  - lexical/morphological lookup: returns multiple analyses based on CURUPIRA lexicon (Martins et al. 2006)
  - disambiguator: finite state disambiguation rules narrow down lexical information, in the spirit of Constraint Grammar (Karlsson et al. 1995; Bick 2000, 2004)
  - parser: bottom-up chart parser establishes relations to check agreement, case and global well-formedness
- Content Analysis:
  - shallow semantic matching strategies between student answer and target, cf. Content Assessment Module (Bailey and Meurers 2006)

Demand-Driven NLP for ICALL
Luiz Amaral & Detmar Meurers
Introduction
Why use NLP
Architectures for using NLP
Research Context
The TAGARELA system
Listening
Reading
Description
Rephrasing
Vocabulary
Fill-in-Blank
System Architecture
NLP analysis modules
Requirements
Annotation-based processing
Activities in the driver's seat
Characteristics of Activities
Relevance for processing
Why each activity matters
Type 1: Text Identity
Type 2: Info. Extraction
Type 3: Inference
Feedback for Type 1
Feedback to be improved
Summary

Our requirements for the NLP architecture

- Allow the analysis manager to flexibly employ NLP modules relevant to a particular activity.
  (→ main topic in the following)
- Flexible control also relevant from NLP perspective, to support interleaving of contributions from modules, e.g.:
  - part-of-speech ambiguity in Portuguese: *a* can be a
    - preposition (*to*)
    - pronoun (*her*, clitic direct object)
    - article (*the*, feminine singular)
    - abbreviation (*association*, *alcoholic*, etc.)
  - tokenization can resolve some part-of-speech ambiguities:
    - *da* = *de* + *a* (article)
    - *vê-la* = *ver* + *a* (clitic pronoun)
    - *à* = *a* (preposition) + *a* (article)
    - *A.A.A.* = *Associação dos Alcólicos Anônimos*

  → TAGARELA tokenizer annotates some part-of-speech

Demand-Driven NLP for ICALL
Luiz Amaral & Detmar Meurers
Introduction
Why use NLP
Architectures for using NLP
Research Context
The TAGARELA system
Listening
Reading
Description
Rephrasing
Vocabulary
Fill-in-Blank
System Architecture
NLP analysis modules
Requirements
Annotation-based processing
Activities in the driver's seat
Characteristics of Activities
Relevance for processing
Why each activity matters
Type 1: Text Identity
Type 2: Info. Extraction
Type 3: Inference
Feedback for Type 1
Feedback to be improved
Summary

Annotation-based processing

- To support a flexible control structure, the data structures serving as input and as output for the analysis modules need to be uniform and explicit.
- NLP analysis = a process of enriching the learner input with annotations (parallel to corpus annotation).
- The same data structure, the learner input annotated with information, is accessed throughout.
  - Closely related idea: Common Analysis System (CAS, Götz and Suhre 2004) of the Unstructured Information Management Architecture (UIMA).
- Which annotation module needs to be called when can be determined automatically by a controller using the input and output specifications of each module.
  - Currently this is hand-coded, but we are planning to realize a demand-driven controller in UIMA.

Demand-Driven NLP for ICALL
Luiz Amaral & Detmar Meurers
Introduction
Why use NLP
Architectures for using NLP
Research Context
The TAGARELA system
Listening
Reading
Description
Rephrasing
Vocabulary
Fill-in-Blank
System Architecture
NLP analysis modules
Requirements
Annotation-based processing
Activities in the driver's seat
Characteristics of Activities
Relevance for processing
Why each activity matters
Type 1: Text Identity
Type 2: Info. Extraction
Type 3: Inference
Feedback for Type 1
Feedback to be improved
Summary

General Characteristics of Activities

Activities can be characterized and differ in:

- task specification, e.g.:
  - listen, read, write, comment, complete
- level, e.g.:
  - basic, intermediate, advanced
- expected input, e.g.:
  - word, phrase, sentence
- nature and availability of target responses and type of variation from target that is permitted
- required skills and abilities, e.g.:
  - strategies needed (e.g., scanning, summarizing, grouping)
  - amount of content manipulation required
  - required awareness of linguistic categories and rules
- pedagogical goals behind activity and feedback provided:
  - generally: improve the required skills and abilities

Where it matters for processing

- General claim: The NLP analysis and feedback generation depend on the specific activity (type).
- The information from the activity model has an impact on
  - **Property Identification**:
    - Which linguistic properties (incl. errors) of the learner input can be observed in a given activity?
  - **Property Selection**: Which of the obtained properties does it select as error cause (or other relevant aspect)?
    - Which of the identified errors should be the focus of the feedback given activity and its specific pedagogical goals?
    - Which of the identified properties is most likely to provide a reliable assessment?
  - **Feedback Strategy**: Which strategy does it chose? E.g.:
    - explicit feedback on form for FIBs
    - scaffolding for reading comprehension (i.e., encouraging the use of required strategies)

Property identification in TAGARELA

- In TAGARELA, different activity types require different linguistic information to analyze student's input:
  - FIB: spell-checking, lexical information
  - Rephrasing: as above + syntactic processing and basic content assessment (correct answer, token matcher)
  - Reading: as above + all content analysis modules
- Why not always run everything?
  - "Don't guess what you know"
  - The more we know the linguistic properties, the types of variation, and the potential errors NLP needs to detect,
    - the more specific information we can diagnose
    - with higher reliability.

From variation between activity types to variation within an activity type

Reading comprehension questions in TAGARELA all take full sentences as answers, yet are heterogeneous in terms of processing and feedback:

- Type 1: Text Identity
  - *student required to:* identify answer in text
  - *content analysis:* string match
- Type 2: Information Extraction
  - *student required to:* extract information from text
  - *content analysis:* token match
- Type 3: Inference
  - *student required to:* deduce information from text
  - *content analysis:* concept (and relation) matching

Type 1: Identify answer in text

Módulos: 1 2 3 4 5 Atividades: 1 2

Leitura

Instrução

Leia o texto e responda às questões usando frases completas e o vocabulário apresentado no texto.

Quem é você?

Meu nome é João de Castro. Eu sou morador do Rio de Janeiro. Tenho trinta anos e estudo medicina na UFRJ.

Meu pai se chama Antônio e minha mãe Júlia. Meu pai é jornalista e trabalha em um grande jornal local chamado "O Globo".

Minha mãe é professora de francês. Além do francês, ela também fala italiano e inglês. Ela trabalha em uma escola particular.

Questões: 1 2 3 4 5 6 7 | Próxima Atividade (2)

Questão 7

Onde trabalha a mãe dele?

Ela trabalha em uma escola particular.

à á â ã é ê í ó ô õ ú ü ç
À Á Â Ã É Ê Í Ó Ô Õ Ú Ü Ç

Enviar

Análise:

Input: **Ela trabalha em uma escola particular.**

That's right!

Report Errors & Suggestions

Type 2: Extract information from text

Regiões do Brasil

O Brasil está política e geograficamente dividido em cinco regiões. Os limites de cada região (Norte, Nordeste, Sudeste, Sul e Centro-Oeste) coincidem sempre com as fronteiras dos estados que as compõem.

A região Norte ocupa a maior parte do território brasileiro, com uma área que corresponde a 45,27% da área total do País. Formada por sete Estados, tem sua área quase totalmente dominada pela bacia do Rio Amazonas.

A região Nordeste pode ser considerada a mais heterogênea do País. Dividida em quatro grandes zonas - meio-norte, zona da mata, agreste e sertão -, ocupa 18,26% do território nacional e tem nove estados.

O Sudeste é formado por quatro Estados. Esta é a região de maior importância econômica do País, onde está concentrado também o maior índice populacional - 42,63% dos brasileiros.

Já o Sul, região mais fria do País, com ocorrências de geadas e neve, é a que apresenta menor área, ocupando 6,75% do território brasileiro e com apenas três Estados. Os rios que cortam sua área formam a bacia do Paraná em quase toda sua totalidade e são de grande importância para o País, principalmente pelo seu potencial hidrelétrico.

Finalmente, a região Centro-Oeste tem sua área dominada basicamente pelo Planalto Central Brasileiro e pode ser dividida em três porções: maciço goiano-mato-grossense, bacia de sedimentação do Paraná e as depressões. Ela é formada por quatro Estados e nela está a capital do Brasil.

Questões: 1 2 3 4 5 6 7 | Próxima Questão (2)

Questão 1

Quantas regiões tem o Brasil?

à á â ã é ê í ó ô õ ú ü ç
À Á Â Ã É Ê Í Ó Ô Õ Ú Ü Ç

Enviar

Análise:

Report Errors & Suggestions

Type 3: Infer information based on text

Módulos: 1 2 3 4 5 Atividades: 1

Leitura

Instrução

Leia o texto e responda as questões usando o vocabulário apresentado pelo texto.

O que você fez na semana passada?

Sou estudante de medicina e estudo na Universidade de São Paulo. Os meus dias são parecidos todas as semanas.

Semana passada eu acordei às 6:30 todos os dias com exceção da terça, quando tive uma aula às 7:00. Por isso acordei às 6:00. Na quarta-feira eu cheguei na universidade às 9:00 e tive duas aulas: anatomia e citologia. Fui almoçar com uns amigos ao meio-dia e depois fui para a minha aula de inglês. Eu estudo inglês duas vezes por semana. Depois do curso eu voltei para a universidade para mais uma aula.

Na quinta à tarde eu fui ao cinema com a minha namorada. Fomos ver um filme americano e depois eu levei ela em casa. À noite, quando cheguei em casa li uns capítulos do meu livro de anatomia para minha prova no dia seguinte.

Questões: 1 2 3 4 5 | Próxima Questão (2)

Questão 1

A que horas ele acordou na segunda da semana passada?

à á â ã é ê í ó ô õ ú ü ç
À Á Â Ã É Ê Í Ó Ô Õ Ú Ü Ç

Enviar

Análise:

Report Errors & Suggestions

Example Feedback for Type 1 Task

Módulos: 1 2 3 4 5 Atividades: 1 2

Leitura

Instrução

Leia o texto e responda às questões usando frases completas e o vocabulário apresentado no texto.

Quem é você?

Meu nome é João de Castro. Eu sou morador do Rio de Janeiro. Tenho trinta anos e estudo medicina na UFRJ.

Meu pai se chama Antônio e minha mãe Júlia. Meu pai é jornalista e trabalha em um grande jornal local chamado "O Globo".

Minha mãe é professora de francês. Além do francês, ela também fala italiano e inglês. Ela trabalha em uma escola particular.

Questões: 1 2 3 4 5 6 7 | Próxima Atividade (2)

Questão 7

Onde trabalha a mãe dele?

Ela trabalha uma escola particular.

à á â ã é ê í ó ô õ ú ü ç
À Á Â Ã É Ê Í Ó Ô Õ Ú Ü Ç

Enviar

Análise:

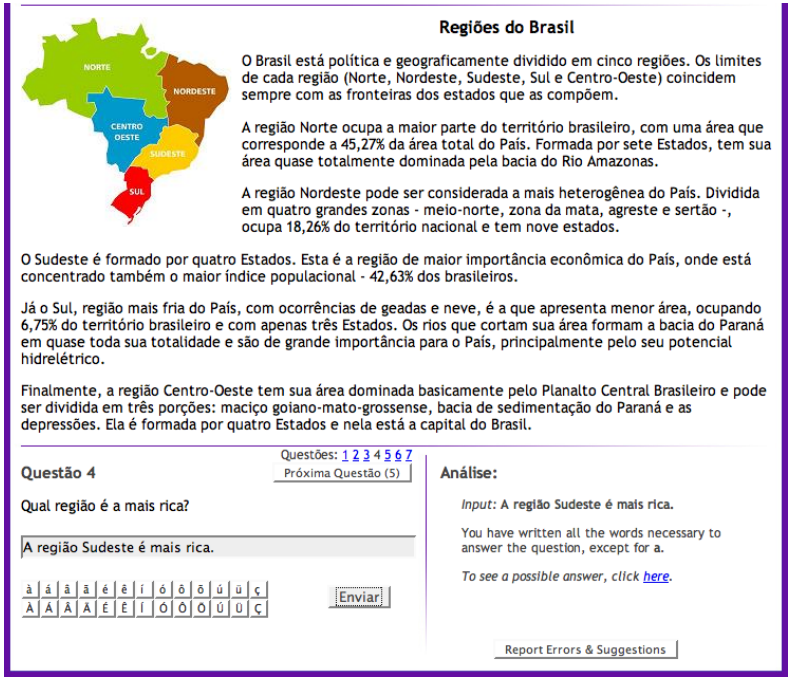
Input: **Ela trabalha uma escola particular.**

You have written all the words necessary to answer the question, except for **em**.

To see a possible answer, click here.

Report Errors & Suggestions

Example Problematic Feedback for Type 3 Task



Summary

- We discussed what we believe is needed to develop ICALL systems which
  - integrate a wider range of activity types,
  - resulting in learner input of a heterogeneous nature,
  - which should be evaluated on different criteria
  - with the feedback also depending on the learner history?
- ICALL architectures would benefit from
  - viewing the NLP modules as enriching the input with annotations, and
  - giving the activity model control over which annotations can or must be provided in a demand-driven architecture
- In the future, we plan to
  - port system to UIMA as a free, robust, large-scale platform
  - derive control automatically from the input-/output specification provided with each NLP module
  - specify more fine-grained activity models supporting finer-grained analyses and feedback distinctions
  - work on a wider range of activity types

References

Amaral, Luiz and Detmar Meurers (2006). Where does ICALL Fit into Foreign Language Teaching? CALICO Conference. May 19, 2006. University of Hawaii.

Amaral, Luiz Alexandre (2004). ICALL Specifications. Ms. Department of Linguistics & Department of Spanish and Portuguese. Ohio State University.

Amaral, Luiz Alexandre and Walt Detmar Meurers (2005). Towards Bridging the Gap between the Needs of Foreign Language Teaching and NLP in ICALL. In António Pedros-Gascon (ed.), *Electronic Proceedings of the 8th Annual Symposium on Hispanic and Luso-Brazilian Literatures, Linguistics, and Cultures*.

Bailey, Stacey and Detmar Meurers (2006). Exercise-driven selection of content matching methodologies. EUROCALL Conference. September 6, 2006. University of Granada.

Bick, Eckhard (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Bick, Eckhard (2004). PaNoLa: Integrating Constraint Grammar and CALL. In Henrik Holmboe (ed.), *Nordic Language Technology, Arbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (Yearbook 2003)*, Copenhagen: Museum Tusculanum, pp. 183–190.

Götz, Thilo and Oliver Suhre (2004). Design and implementation of the UIMA Common Analysis System. *IBM Systems Journal* 43(3), 476–489.

Heift, Trude and Devlan Nicholson (2001). Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education* 12(4), 310–325. http://aied.inf.ed.ac.uk/members01/archive/vol_12/heift/paper.pdf.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä and Arto Anttila (eds.) (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. No. 4 in Natural Language Processing. Berlin and New York: Mouton de Gruyter.

Martins, Ronaldo, Ricardo Hasegawa and Maria das Graças Nunes (2006). Curupira: a functional parser for Brazilian Portuguese. In *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR. Lecture Notes in Computer Science 2721*. Faro, Portugal: Springer.