Will VOICE go POS?

Ruth Osimk, University of Vienna

VOICE — Vienna Oxford International Corpus of English

universität wien

ABSTRACT

After the release of VOICE 1.0 Online a year ago, the main focus in the next stage of the project is on increasing the applicability and usability of the corpus. One possibility is extending the corpus mark-up by **adding part-of-speech (POS) tags** to indicate word class categories. Unsurprisingly, this proves to be a highly challenging task when applied to a corpus of spoken ELF. With no precedence of POS taggers applied to ELF, there is no other choice but to explore the suitability of existing taggers, and the possibilities of adapting these to meet the specific requirements of VOICE data.

This VP will focus on two particularly prominent issues involved in assigning POS tags to VOICE data, namely the classification of unconventional items and the relationship between forms and functions in ELF (e.g. Seidlhofer 2009). It will discuss how these features have been dealt with in existing L2 corpora, focusing particularly on the method of 'error-free' SLA-tagging using *TreeTagger* (Rastelli 2009). As a result of comparing more traditional methods and SLA-tagging, this VP will suggest what might be a suitable tool for tagging the ELF data in VOICE, and will consider methodological and practical implications of the tagging process for ELF research.

CONTACT

Ruth Osimk
University of Vienna
Email: ruth.osimk@univie.ac.at
Website: www.univie.ac.at/voice

INTRODUCTION: POS TAGGING AN ELF CORPUS

What is POS Tagging?

• "enriching a corpus by adding a part-of-speech tag to each word" (Atwell 2008: 501), incl. inflectional information and lexicosemantic properties (Voutilainen 1999 :5)
• A **pre-requisite** for other types of annotation (e.g. parsing, semantic tagging, lemmatization) (Leech & Smith 1999: 26 ff.)

Why POS tag VOICE?

• to **enrich** the corpus *mark-up*
• to **enhance** the *usability* of the corpus further
• to **enable** *further insights on ELF*, e.g. on lexicogrammatical features and form-function relationships

CHALLENGES FOR POS TAGGING VOICE

1) **Challenge of spoken data** (incomplete starts, repetition, question of chunking, specific mark-up, etc.)
2) **Challenge of ELF data** (code-switching, non-codified items, ENL form/function match might not be applicable, etc.)

Non-codified items (cf. Widdowson 1997; Pitzl, Breiteneder & Klimpfinger 2008)

• S1: i was just like (.) <pvc> **putted** </pvc> things in (LEcon565:378, L1=it-IT)
• S1: it was like a surreal <pvc> **inscenation** </pvc> or something (LEcon573:76; L1=ger-DE)
• S1: with a diverse <pvc> **linguistical** </pvc> (.) group. (EDwsd303:387, L1=dut-NL)

Non-ENL form/function mapping

• S2: you can TASTE it it **taste** even of milk but it's FINE. (LEcon566:183, L1=it-IT)
• S1: i don't want erm let's say this way i also didn't **spoke** to [first name34] in france (PBmtg27:527, L1=ger-DE)
• S2: one can apply (at) this italian agency for (.) two **month** (.) (PRcon550:30, L1=slv-SI)
• S1: in [org11] video camera has **break** down (PBmtg27:653, L1=ger-DE)
• S1: the waiter is NOT somebody who **hate** americans (EDsed31:964, L1=ger-AT)
• S11: but er (1) we are <3> **complete** different </3> (EDsed31:1134, L1=it-IT)

PILOT STUDY: TEST-TAGGING VOICE – Example: <pvc>s

A) STUDY DESIGN

4 speech events
• LEcon565, LEcon566, LEcon573, LEcon575
• 2 speakers (L1 Italian, L1 German)

Chunking according to utterances
• all utterances containing a <pvc> → automatic retrieval of utterances from corpus mark-up
• Total no. of utterances: **20**
• Total no. of tokens: **543**
• Total no. of <pvcs>: **22**

B) METHOD

Retrieval of all utterances <u> containing a <pvc>

→ Tagging of utterances with **TreeTagger (in GATE)**

→ Control: **Manual tagging** of retrieved utterances

→ **Comparison** of tagging results

C) RESULTS

Comparison manual tagging vs. **TreeTagger:**
Overall accuracy of full utterances containing <pvc>: **84,5%**
Accuracy of <pvc>s: **13 agree**/**9 disagree**
Without 10 ignored words and <pvc>s: **87,5 %**

Pvc	Word with immediate co-text	Manual Tag	Tag TreeTagger
1	is it **spanishy** or	JJ	JJ 0.929969
2	or **portuguesey** whatever shop	JJ	NN 0.730337
3	you say **anachrom**	NN	NN 0.532445
4	just like **putted** things in	VVD	VVN 0.989772
5	is slightly **liquidy** but	JJ	JJ 0.794834
6	is more **liquidy** yeah	JJ	NN 0.889958
7	it isn't **liquidy** i think	JJ	JJ 0.914492
8	never then **chinesey** ones	JJ	JJ 0.905536
9	have something **liquidy** then	JJ	NN 0.975937
10	not really **softish** huh	JJ	JJ 1.000000
11	just like **claustrophobicy** get	JJ	NN 0.987108
12	they look all **frenchers**	NNS	NNS 0.918265
13	but **slutty**	JJ	JJ 1.000000
14	and those **slutty** forty year	JJ	JJ 1.000000
15	a surreal **inscenation** or something	NN	NN 0.996255
16	like coordinate **subcontractors**	NNS	NNS 1.000000
17	grey zone **anyways**	RB	RB 1.000000
18	are not **dimensioned** we can't	JJ	VVD 0.501130
19	did not **re-enrol** he probably	VV	NN 0.682976
20	students don't **re-enrol** for	VV	NN 0.809839
21	you didn't **re-enrol**	VV	NN 0.510386
22	create the **sotteck** socket	NN	NN 0.720744

TAGGING PRACTICES – APPLICABLE TO VOICE?

	Speaker is viewed as...	Non-codified items	Mapping of forms and functions	GOALS
Error-tagging	Language learner	Errors → error tag is attached	according to assumed TL goal e.g. *he jump 3rd person singular	Investigate language of learners • to gain new insights for language pedagogy and SLA • by using methods from SLA which compare learner language to native standards (CA, CEA)
SLA-tagging	L2 user (though TL is goal)	Manifestations of interlanguage → 'virtual categories'	according to form → kept separate in order to investigate the interlanguage system & its mechanisms	Investigate • systematicity of form-function mapping and development of TL categories with learners • reveal unexpected features of the learner language
ELF-tagging	Language user (TL not necessarily goal)	Manifestations of virtual language → **tagging?**	**?**	Development of tagging system which is meaningful for ELF research → e.g. forms/function use for communicative purposes

Fig 1: A sample decision tree (Schmid 1994)

TreeTagger (Schmid 1994)
• **Statistical tagger**, based on L1
• Assigns tags according to **lexical root, morphology** and **context**, using a **decision tree**
• Outputs high/low **confidence rates**

NEXT STEPS...

Non-codified items (<pvc> s)	• feed into lexicon
Non-ENL form/function mapping	• tag non-codified items according to form or function?

REFERENCES

• Atwell, Eric. 2008. "Development of tag sets for part-of-speech tagging". In Lüdeling, Anke; Kytö, Merja (eds.). *Corpus linguistics: An international handbook*. Berlin/New York: Mouton de Gruyer, 501-527.
• Leech, Geoffrey; Smith, Nicholas. 1999. "The use of tagging". In van Halteren, Hans (ed.). *Syntactic wordclass tagging*. Dordrecht: Kluwer, 23-36.
• Pitzl, Marie-Luise; Breiteneder, Angelika; Klimpfinger Theresa. 2008. "A world of words: processes of lexical innovation in VOICE". *Views* 17, 21-46.
• Rastelli, Stefano. 2009. "Learner corpora without error tagging", *Linguistik online* 38, 57-66.
• Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*.
• Seidlhofer, B. 2009. "Orientations in ELF research: form and function". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: Studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 37-59.
• VOICE. 2009. *The Vienna-Oxford International Corpus of English*. (version 1.0 online) http://voice.univie.ac.at (18 May 2010)
• Voutilainen, Atro. 1999. "Orientation". In van Halteren, Hans (ed.). *Syntactic wordclass tagging*. Dordrecht: Kluwer, 3-8.
• Widdowson, H. 1997. "EIL, ESL, EFL: global issues and local interests". *World Englishes* 16, 135-146.