ICE-Scotland

Release 01 – first parts

April 2020

1. Project team

The compilation of ICE-Scotland was directed by Ulrike Gut (gut@wwu.de) at the University of Münster in collaboration with Ole Schützler at the University of Bamberg. Project members were Jenny Herzky, Zeyu Li, Gavin Andrews, Sin Yu Bonnie Ho, Lena Kastner, Anna Konstantinova, Alice Limmer, Philipp Meer, Alina Pepper, Nicholas Peterson, Johannes Trüdinger, Andreas Weilinghoff and Roman Zingel. We gratefully acknowledge the generous assistance of Manfred Krug in Bamberg as well as our partners at Glasgow University (especially Jennifer Smith). We are grateful to the Scottish Parliament, STV, various Scottish newspapers and radio stations and everyone else who gave us permission to use their data.

2. Corpus format

The written part of the ICE-Scotland is available either as txt files, as xml files or in a POS tagged version (see section 3). The spoken part is available as ELAN eaf files (xml files) and as txt files. In addition, we are making all of the raw files available. We did not follow the ICE convention of arranging the corpus data in texts of 2,000 words each, but constructed separate files for each text produced by one author. The required number of words per text category adheres to the ICE guidelines (see section 4).

The following metadata is included (if available) in each xml file: transcriber's initials, place and time of text production, author's gender, age and regional background. The symbol '?' is used when metadata was not available. The spreadsheet "metadata_ICE_Scotland.xlsx" lists all speakers and writers represented in ICE-Scotland with their gender, age, regional background, and occupation. Please note that some speakers appear in several files, for example some journalists who appear in different broadcast news files and politicians who appear in the parliamentary debates and possible also a broadcast discussion.

3. Annotation of the corpus

ICE-Scotland was annotated using **Pacx**, the **P**latform for **A**nnotation of **C**orpora in **X**ML (http://pacx.sourceforge.net/), developed by Holger Voormann.

3.1 Spoken part

The spoken data was **transcribed orthographically** and **time-aligned** using ELAN (https://tla.mpi.nl/tools/tla-tools/elan/download/). Time-aligned **phonemic transcriptions** were created using WebMAUS (Schiel 2004) and were corrected manually. The annotation process is described in Schützler et al. (2017), Gut & Fuchs (2017) and Gut (2011). Figure 1 shows a part of the transcription of file bint_04, a broadcast interview.

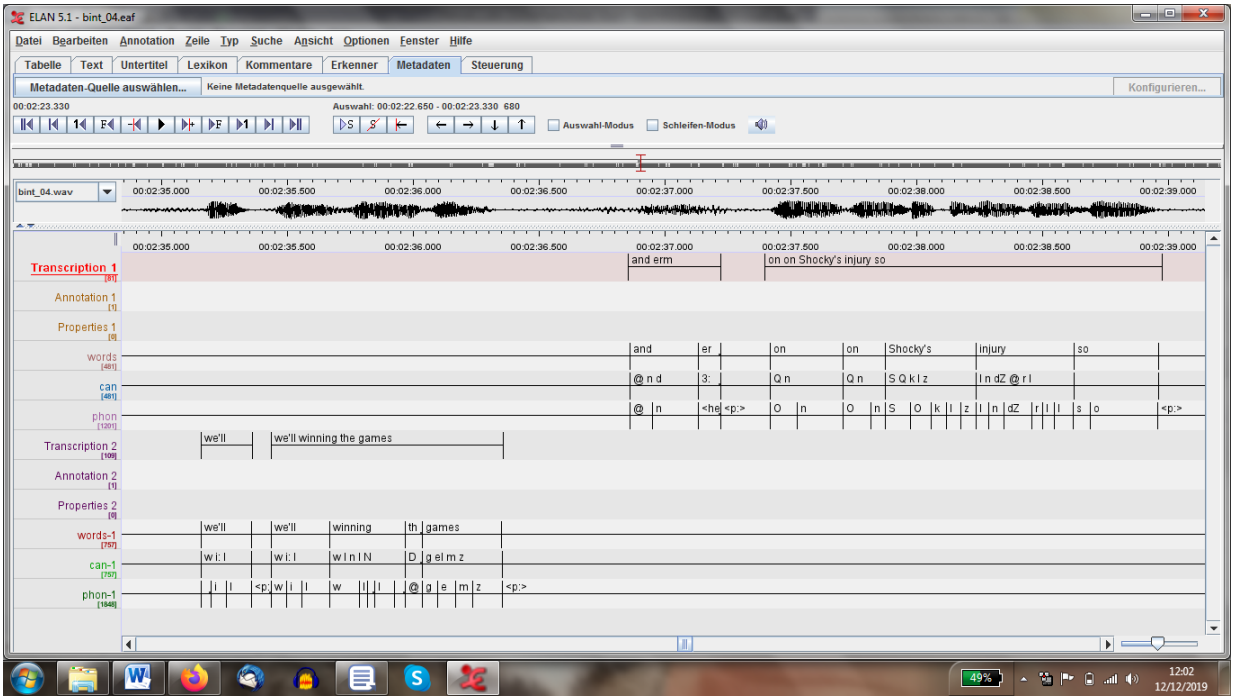

***Figure 1**. Transcription of the spoken data in ICE-Scotland.*

For each speaker, the orthographic transcription of what was said is provided on the tier called 'Transcription X', divided into intonation phrases. Below this, on the **'Annotation' and 'Properties' tiers**, further annotations may be given such as:

Other language	<other-language>
anonymised word/s	<anonymous>
unclear word/s	<unclear>
interrupted utterance	<interrupted>
hesitation	<hesitation>
repaired utterance	<repair>
quoted word/s	<mention>
transcriber's comment	<transcriber's comment>

On the **'words' tier**, the boundaries of all words and their orthographic transcription is given. On the **'can' tier** the canonical phonemic transcription created by WebMAUS for this word is given. Please note that this has not been corrected manually. On the **'phon' tier**, the boundaries of all phonemes as well as their phonemic transcription in SAMPA are given. Only phonemes that were actually pronounced are transcribed, i.e. if a speaker produced the word *friend* as [fren], only those four phonemes are transcribed. The phonemic transcription was carried out following the description of Scottish Standard English phonology by Stuart-Smith (2008). The SAMPA symbols for vowels were used as shown in Table 1:

SAMPA	*lexical set*	*examples*	*IPA*
i	FLEECE	*beat, leave*	i
I	KIT + BIRTH	*him, tic + girl*	ɪ, ɜ
e	FACE + SQUARE	*fail, tame + hair, bear*	eɪ, e
E	DRESS + BERTH	*mess, bet + kerb, herb*	ɛ, ɜ
a	TRAP + PALM + BATH	*hat, can + father, calm*	a, æ
U	GOOSE + FOOT	*lose, fool + put, took*	ʊ, u
o	GOAT	*rode, told*	əʊ
O	LOT + THOUGH	*rob, lock + fall, taught*	ɒ, ɔ
V	STRUT + NURSE + COMMA	*but, luck + further, hurt, word + sofa*	ʌ, ɜ
@	LETTER and function words	*arise, father + for, and, of…*	ə
Vi	PRICE	*file, bike*	aɪ
Oe	CHOICE	*boil, coin*	ɔɪ
VU	MOUTH	*house, loud*	aʊ

Table 1. *SAMPA symbols used for the phonemic transcription of the vowels.*

The SAMPA symbols used for the consonants are listed in Table 2.

SAMPA	*Examples*	*IPA*
p	***p****en, ti****p***	p
b	***b****ut, we****b***	b
t	***t****wo, s****t****ing, be****t***	t
d	***d****o, o****dd***	d
tS	***ch****air, na****t****ure, tea****ch***	tʃ
dZ	***g****in, ****j****oy, e****dge***	dʒ
k	***c****at, ****k****ill, s****k****in, ****q****ueen, thi****ck***	k
g	***g****o, ****g****et, be****g***	g
f	***f****ool, enou****gh****, lea****f***	f
v	***v****oice, ha****ve****, o****f***	v
T	***th****ing, brea****th***	θ
D	***th****is, brea****the***	ð
s	***s****ee, ****c****ity, pa****ss***	s
z	***z****oo, ro****s****e*	z
S	***sh****e, ****s****ure, emo****ti****on, lea****sh***	ʃ
Z	*plea****s****ure, bei****ge***	ʒ
h	***h****am*	h
m	***m****an, ha****m***	m
n	***n****o, ti****n***	n
N	*si****ng****er, ri****ng***	ŋ
l	***l****eft, be****ll***	l
r	***r****un, ve****r****y*	ɹ
w	***w****e*	w
j	***y****es*	j
x	*lo****ch***	x
C	*drei****ch***	ç

Table 2. *SAMPA symbols used for the phonemic transcriptions of consonants in ICE Scotland.*

In general, the **'phon' tier** contains **phonemic transcriptions**, i.e. even if a FACE vowel is realised as a diphthong by a particular speaker in a particular word, it is transcribed as 'e'. There are only three features where **phonetic transcriptions** were provided, i.e. where the actual pronunciation of a sound was transcribed. This was the case for the realisation of a word with <wh-> as in *what*, the realisation of word-medial /t/ as in *network* and the realisation of coda /r/ as in *car*. As Table 3 shows, for these three cases it was transcribed whether [ʍ] or [w] was produced, whether the word-medial /t/ was glottalised or not and whether coda /r/ was realised or not. If coda /r/ was realised, a '~' was added to the preceding vowel, no matter whether coda /r/ was realised as a tap, trill, approximant or any other rhotic sound.

Phonetic transcriptions	
what	ʍ or w
ne**t**work	ʔ or t
ca**r**	a~ or a

Table 3. *Phonetic transcription of realisation of <wh->, word-medial t and coda /r/.*

Silent pauses were marked as '<p:>' and both filled pauses ('erm', 'mhm' and so on) and mispronunciations were marked as '<hes>'.

3.2 Written part

The written data was transcribed using *Vex*. We did not use the SGML tags described in the ICE manual (http://ice-corpora.net/ice/manuals.htm), but used xml tags instead. The following metadata and text features were annotated with the following tags (see also "AnnotationSchema.xml"):

Metadata

metadata	<meta>
Transcriber	<transcriber>
Date when text was produced	<date>
Place where text was produced/published	<place>
Author's gender	<author><gender>
Author's age	<age>
Author's regional background	<regional background>

Text features written part

Heading	<heading>
Paragraph	<p>
Unclear word/s	<unclear words> e.g. <unclear words="1"/>
Line break	<break />
Unusual punctuation	<punctuation>
Word in boldface	<boldface>

Word in italics	<italics>
Underlined word	<underline>
Superscript	<superscript>
Subscript	<subscript>
Small capitals	<small-capitals>
Change of typeface	<typeface-change>
Quotation	<quote>
Indigenous word	<indigenous>
Colloquialism	<coll> e.g.: 4 you <coll for="for you">
Deleted word/s	<deleted>
Discontinuous word	<discontinuous>
Signature	<signature>
Footnote	<footnote>
Untranscribable elements (graphs, formulas…)	<object> e.g. <object type="graphic"/>
Spelling error	<correction> e.g. <correction original="moretimely"><w c7="RGR">more</w><w c7="RR">timely</w></correction>
Superfluous words	<error superfluous=”yes”>is</error>
Incomplete words (SPOKEN)	<error>ga-</error>
Mentions	<mention>
Editorial comment	<editor-comment>
Other language	<other-language>
Anonymised names	<x-anonym-x> e.g. <x-anonym-x type="family-name"/>
Extra-corpus material	<unannotated>

3.3 POS-tagging (written & spoken part)

Automated POS-tagging was applied to the entire corpus by using the *ClawsAnt* Tagger and the CLAWS7 tagset (http://ucrel.lancs.ac.uk/claws7tags.html). POS-tags were then manually corrected and/or modified where necessary. The most important tagging procedures to note are the following, as they might deviate from previous or ‘standard’ procedures:

XML structure

The xml file structure goes down to the individual word level. POS-tags are specified as “c7 =” (cf. CLAWS7) rather than “pos =”, e.g.

<w c7="NN1">whisky</w>

- **Important difference xml vs. txt files:**

Note that the xml version contains the **text in its original version**, i.e. including any errors. Such errors have been error-tagged if they are not part of the area of grammar:

<w c7="II">in</w><w c7="JJ">common</w>**<correction original="closes"><w c7="NN2">closets</w></correction>**<w c7="CC">and</w><w c7="NN2">stairwells</w>

In the txt version, however, **only the corrected version** has been kept, without any error-markup:

in_II common_JJ **closets_NN2** and_CC stairwells_NN2

Punctuation marks

The CLAWS7 tagset does not contain a specific category for punctuation marks. As a consequence, the 'default' tag assigned to any punctuation or similar characters was kept, i.e., they were tagged 'as themselves':

._.	/_/
,_,	(_(
?_?	[_[
|_|	etc.

but:

"_" (**"** in the XML-version)
&_CC (coordinating conjunction, like *and*; **&** in the XML-version)

Compounds and compound-like constructions

Word classes were kept in 'fixed expressions'. Capitalized compounds and proper nouns were usually tagged 'NP', rather than just 'NN'. Compounds were not ditto-tagged[1].

General Elections – General_JJ Elections_NP2
the Second World War – the_AT Second_MD World_NP1 War_NP1
the Culloden campaign – the_AT Culloden_NP1 campaign_NN1
school teachers – school_NN1 teachers_NN2, **not:** school_NN121 teachers_NN122

Contractions and possessives

All contractions as well as instances involving apostrophes (e.g. possessives) were split into separate entities, e.g.

Scotland's mountains – Scotland_NP1 's_GE mountains
We can't wait – We ca_VM n't_XX wait

[1] So-called "ditto tags" mark strongly-bound strings of words, e.g. *in terms of* – in_II31 terms_II32 of_II33; *once and for all* – once_RR41 and_RR42 for_RR43 all_RR44.
According to the Lancaster University CLAWS7 website, a ditto tag "occurs as part of a sequence of similar tags, representing a sequence of words which for grammatical purposes are treated as a single unit" (http://ucrel.lancs.ac.uk/claws7tags.html).

We'll see – We_PPIS2 'll_VM see
I'm here – I_PPIS1 'm_VBM here_RL
They're here – They_PPHS2 're_VBR here_RL
He's been searching – He_PPHS1 's_VHZ been_VBN searching_VVG

Furthermore, please note that the files belonging to the **POS-tagged spoken part** of the corpus have been split into **separate tiers if multiple speakers were involved**. Each speaker received a separate "sub-file", e.g. bdis_01_SP1, bdis_01_SP2, etc.

4. Word count of text categories

Text type (ICE text category label)	***ICE-Scotland file name***	***Words***
Academic writing humanities (W2A)	AHum_01 – 10	20,144
Academic writing natural sciences (W2A)	ANat_01 – 10	20,221
Press reportage (W2C)	Rep_01 – 79	40,264
Instructive writing/skills and hobbies (W2D)	SkHo_01 – 27	20,356
Broadcast news (S2B)	bnew_01 – 37	40,150
Broadcast talks (S2B)	btal_01 – 50	40,265

References

Gut, Ulrike. (2011). Language documentation and archiving with Pacx, an XML-based tool for corpus creation and management. In Nathan David (ed.), *Proceedings of the Workshop on Language Documentation and Archiving, London*, pp. 21-25.

Schiel, Florian. 2004. MAUS goes iterative. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lissabon, Vol. 3, Maria Teresa Lino (ed.), 1015–1018. Paris.

Schützler, O., Gut, U. & Fuchs, R. (2017). New perspectives on Scottish Standard English. Introducing the Scottish component of the International Corpus of English. In Hancil, S., Beal, J. (Eds.), *Perspectives on Northern Englishes* (p. 273-301). Berlin: Mouton de Gruyter.

Stuart-Smith, J. (2008). Scottish English: Phonology. In: Bernd Kortmann & Clive Upton, (eds.), *Varieties of English*. Vol. 1: The British Isles. Berlin: Mouton de Gruyter. 48-70.

5. Publications based on ICE-Scotland

Gut, U. & Fuchs, R. (2017). Exploring speaker fluency with phonologically annotated ICE corpora. World Englishes, 36(3), 387-403.Gut, Ulrike and Fuchs, Robert. (2013). Progressive aspect in Nigerian English. *Journal of English Linguistics* 41(3), 243-267.

Schützler, O., Gut, U. & Fuchs, R. (2017). New perspectives on Scottish Standard English. Introducing the Scottish component of the International Corpus of English. In Hancil, S., Beal, J. (Eds.), *Perspectives on Northern Englishes* (p. 273-301). Berlin: Mouton de Gruyter.