

# ICE-Scotland

Release 02

July 2022

ICE-Scotland by Prof. Dr. Ulrike Gut and Dr. Ole Schützler is licensed under a [Creative Commons Attribution-Non-Commercial-Share Alike 3.0 Germany License](#). Permissions beyond the scope of this license may be available upon request.



## 1. Project team

The compilation of ICE-Scotland was directed by Ulrike Gut ([gut@wwu.de](mailto:gut@wwu.de)) at the University of Münster in collaboration with Ole Schützler at the University of Leipzig ([ole.schuetzler@uni-leipzig.de](mailto:ole.schuetzler@uni-leipzig.de)). 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).

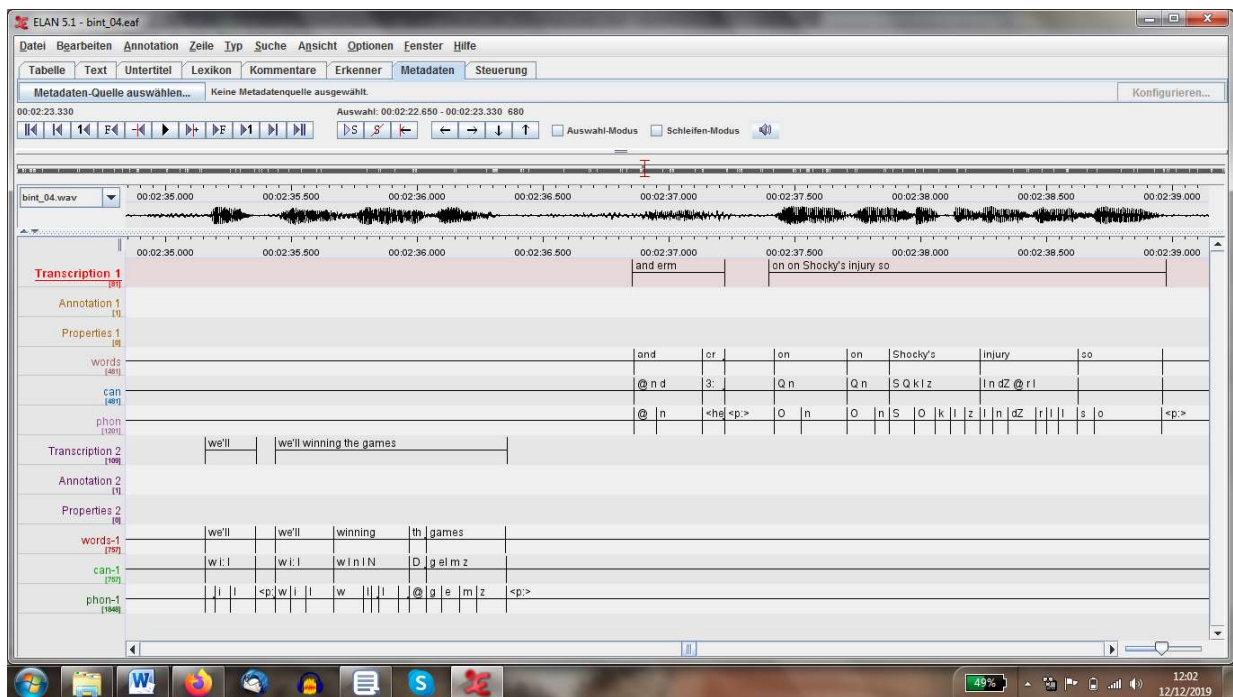
The following metadata is included (if available) in each xml file: transcriber's initials, place and time of text production, author's gender, age and regional background. The symbol '?' is used when metadata was not available. The spreadsheet "metadata\_ICE\_Scotland.xlsx" lists all speakers and writers 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 possibly also a broadcast discussion.

## 3. Annotation of the corpus

ICE-Scotland was annotated using **Pacx**, the **Platform for Annotation of Corpora in XML** (<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:

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

<b>SAMPA</b>	<b>Examples</b>	<b>IPA</b>
<b>p</b>	<i>pen, tip</i>	p
<b>b</b>	<i>but, web</i>	b
<b>t</b>	<i>two, sting, bet</i>	t
<b>d</b>	<i>do, odd</i>	d
<b>tS</b>	<i>chair, nature, teach</i>	tʃ
<b>dZ</b>	<i>gin, joy, edge</i>	dʒ
<b>k</b>	<i>cat, kill, skin, queen, thick</i>	k
<b>g</b>	<i>go, get, beg</i>	g
<b>f</b>	<i>fool, enough, leaf</i>	f
<b>v</b>	<i>voice, have, of</i>	v
<b>T</b>	<i>thing, breath</i>	θ
<b>D</b>	<i>this, breathe</i>	ð
<b>s</b>	<i>see, city, pass</i>	s
<b>z</b>	<i>zoo, rose</i>	z
<b>S</b>	<i>she, sure, emotion, leash</i>	ʃ
<b>Z</b>	<i>pleasure, beige</i>	ʒ
<b>h</b>	<i>ham</i>	h
<b>m</b>	<i>man, ham</i>	m
<b>n</b>	<i>no, tin</i>	n
<b>N</b>	<i>singer, ring</i>	ŋ
<b>l</b>	<i>left, bell</i>	l
<b>r</b>	<i>run, very</i>	ɹ
<b>w</b>	<i>we</i>	w
<b>j</b>	<i>yes</i>	j
<b>x</b>	<i>loch</i>	x
<b>C</b>	<i>dreich</i>	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 [ʌ] 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.

<i>Phonetic transcriptions</i>	
<b>what</b>	ʌ or w
<b>network</b>	? or t
<b>car</b>	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:**

“\_” (&quot; in the XML-version)

&\_CC (coordinating conjunction, like *and*; &amp; 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<sup>1</sup>.

*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

*We’ll see* – We\_PPIS2 ’ll\_VM see

<sup>1</sup> 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/claws7/tags.html>).

*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

<i>Text type (ICE text category label)</i>	<i>ICE-Scotland file name</i>	<i>Words</i>
Academic writing humanities (W2A)	AHum_01 – 10	20,144
Academic writing natural sciences (W2A)	ANat_01 – 10	20,221
Academic writing social sciences (W2A)	ASoc_01-10***	20,106
Academic writing technology (W2A)	ATec_01-11	20,073
Editorials (W2E)	Ed_01-32***	20,486
Essays (W1A)	Es_01-11***	20,130
Popular writing humanities (W2B)	PHum_01-27***	20,197
Popular writing natural sciences (W2B)	PNat_01-41***	20,039
Popular writing social sciences (W2B)	PSoc_01-29***	20,273
Popular writing technology (W2B)	PTec_01-30***	20,051
Press reportage (W2C)	Rep_01 – 79	40,264
Instructive writing/skills and hobbies (W2D)	SkHo_01 – 27	20,356
Broadcast discussions (S1B)	bdis_01-21*	40,170
Broadcast interviews (S1B)	bint_01-15	20,136
Broadcast news (S2B)	bnew_01 – 37	40,150
Broadcast talks (S2B)	btal_01 – 50	40,265
Legal presentations (S2A)	leg_01-26	20,576
Non-broadcast talks (S2B)	nbтал_01-15**	20,024
Parliamentary debates (S1B)	parl_01-06	20,658

\* The files bdis\_04, 05, 12, 15, 17, 18 and 19 do not contain any phonemic annotations

\*\* The files nbтал\_01, 02, 03, and 15 do not contain any phonemic annotations

\*\*\* The files ASoc\_07–10; Ed\_01–07; Es\_01–11; PHum\_12–27; PNat\_01, 03–05, 07, 09, 26–41; PSoc\_03, 13–29; PTec\_23–30 do not contain any POS-tagging



## 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.