Constructing E-Language Corpora: a focus on CorCenCC (The National Corpus of Contemporary Welsh)

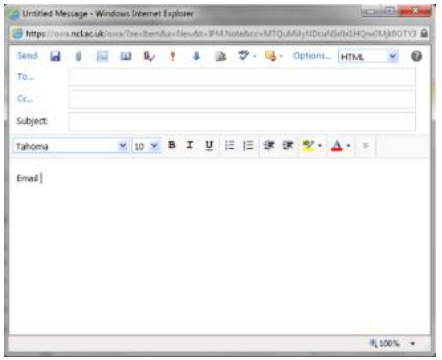
Dawn Knight, Cardiff University, Wales, UK

Overview

1. Definitions and context
2. CANELC – mapping the 'value' of e-language corpora
3. CorCenCC
4. Corpus design and construction - methodological, technical and practical issues and challenges
   - Planning and piloting; sampling; (meta)data extraction and anonymisation; classification/tagging visualisation and analysis – constructing corpus infrastructure
5. Ethical considerations
6. Current progress/closing remarks

1. Definitions and context

- E-language = any communicative, interactive and/or linguistic stimulus that is digitally based and 'incorporates multiple forms of media bridging the physical and digital' (Boyd & Heer 2006: 1).



1. Definitions and context

- An increasing amount of corpora are starting to include e-language in their design but, to date, the majority of work in corpus linguistics on the description of e-language has focused on using either small-scale or bespoke corpora.

- Few corpora in existence which allow users to comment on e-language use in general. This has meant that the ways in which we live and communicate in the digital world 'across multiple resources, remains an under-explored area of research in corpus linguistics' (Knight et al., 2013: 30).

2. CANELC

- **CANELC** = The Cambridge and Nottingham E-language Corpus
- Contains data from 2010-2011. Built in 2011.
- CANELC aimed to include contributions:
  - from a range of different sociolinguistically profiled participants
  - With a word count divided equally among the different 'types' of data

Data Type	Number of Contributors	Number of Messages/ Entries	Word Count	
			Raw	%
Twitter	30	18972	259101	26%
Blogs	36	1101	267983	27%
Discussion Boards	12	2715	242727	24%
Emails	Various	1920	128951	13%
SMS	11	5215	101913	10%
		29923	**1000675**	**100%**

2. CANELC

Variable	Aimed composition	Actual composition
Number of participants	10–40 per source	11–36 contributors per source
Gender	50:50 male and female	50 percent of the corpus has a circa 50:50 balance. For 50 percent genders are unknown
Age	Under 19, 10 percent of total 20–24, 10 percent of total 25–29, 10 percent of total 30–34, 10 percent of total 35–39, 10 percent of total 40–44, 10 percent of total 45–49, 10 percent of total 50–54, 10 percent of total 55–59, 10 percent of total Over 60, 10 percent of total	Contributors were from a range of different age groups although the most populous groupings were 20–24 and 25–29 (there was not a strict balance of contributions across the groupings)
Time period	Contributions posted from 2006–2011	Data from each contributors was collected over a minimum of three days, the majority within the 2010–2011 period
Location	100 percent posting to sites ending in .co.uk	All sites ended in .co.uk and most contributors identified themselves as being British

2. CANELC: initial findings

- The use of **personal pronouns**; adverbs; verbs and interjections is characteristic of more informal communication. Nouns, adjectives, prepositions and articles are more frequent in more 'formal' types of language Heylighen and Dewaele (2003).
- **Modality:** *Could* and *would* are particularly characteristic of spoken, informal discourse, fiction and interpersonal encounters while in more formal, transactional encounters the use of modal verbs is reportedly less frequent (Farr et al., 2004: 13).
- **Hedging:** Hedges are 'expression[s] of tentativeness and possibility' (Hyland, 1996: 433) which operate to 'mitigate the directness of what we say and so operate as face-saving devices' (O'Keeffe et al., 2007: 174).

2. CANELC: initial findings

- **Pronouns and deictic markers:** the rate of use in discussion boards, SMSs and emails mirrors that of spoken discourse, blogs and tweets of written.

- **Modality:** the rate of use in SMSs and discussion boards and emails mirrors that of spoken discourse, tweets and blogs of written.

- **Hedging:** the rate of use in SMSs and discussion boards mirrors that of spoken discourse, blogs, emails and tweets of written.

2. CANELC: initial findings

- Despite being near-immediate, highly interpersonal and semi-synchronous, e-language lacks the utility for effectively communicating 'beyond the word'. In f2f interaction we can access a variety of gestural, paralinguistic and extra-linguistic cues which work with spoken language to generate meaning.

- While contextual cues and emoticons help with this (see Park et al., 2014), we are more reliant on **what** is being said rather than **how** it is said in e-language. We rely on the language alone to build and maintain relationships; to ensure that discourse is polite and non-face-threating, making linguistic devices that function in an interpersonal way.

3. CorCenCC: what is it?

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- **CorCenCC**: Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh: A community driven approach to linguistic corpus construction

- Open-access and freely available 10 million word corpus of Welsh language

- Inter-disciplinary – Computer Science, Applied Linguistics and Education

- Initial conception in November 2011. £1.8m ESRC and AHRC funding obtained in 2015

3. CorCenCC: what is it?

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

"UNESCO Atlas of the world's languages in danger"

- vulnerable
- definitely endangered
- severely endangered
- critically endangered
- extinct

Vulnerable = "most children speak the language, but it may be restricted to certain domains (e.g., home)"

3. CorCenCC: what is it?

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- Extensive community interest in sustaining and 'growing' Welsh
  - largest bilingual community in the UK
  - 20% population of Wales are users of Welsh
  - talking about language, as well as using language to talk, is a feature of Welsh speakers' repertoire

- A rich environment for a resource that focuses on language description rather than prescription.
  - Not always straightforward – linguistic purism is often encountered in Wales

3. CorCenCC: what is it?

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- **Balanced** re. communication type (spoken, written, e-language), genre, language variety (regional, social), thematic context.
- **Representative** of the 562,000 speakers of Welsh in Wales
  - Age
  - Gender
  - Occupation
  - Location
  - Language variety
  - Social and educational backgrounds
- **Representative** of the language use of those speakers
  - i.e. the types of texts that Welsh speakers produce/receive

3. CorCenCC: innovation

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Written

Medium	%	Geiriau / *Words*
Llyfrau / Books	41.75%	1,670,000
Cylchgronau, Papurau Newydd, Cyfnodolion / Magazines, Newspapers, Journals	19.25%	770,000
Deunydd heb ei gyhoeddi / *Unpublished material*	39%	1,560,000
	100%	**4,000,000**

Spoken

Context	%	Geiriau / *Words*
Cyhoeddus/Sefydliadol / *Public/Institutional*	12.5%	500,000
Cyfryngau / *Media*	15%	600,000
Trafodol / *Transactional*	12.5%	500,000
Proffesiynol / *Professional*	10%	400,000
Pedagogaidd / *Pedagogical*	10%	400,000
Cymdeithasu / *Socialising*	20%	800,000
Preifat / *Intimate*	20%	800,000
	100%	**4,000,000**

Based on previous corpora inc. BNC, CANELC and CANCODE

E-Language

Thema/Pwnc / *Theme/Topic*	%	Geiriau / *Words*
Blog	30%	600,000
Gwefan / *Website*	30%	600,000
Ebost / *Email*	20%	400,000
SMS	20%	400,000
	100%	**2,000,000**

3. CorCenCC: team

- CorCenCC Management Team
  - Dawn Knight (PI), Applied/Corpus Linguist
  - Tess Fitzpatrick (CI), Applied Linguist
  - Steve Morris (CI), Welsh Language expert

Swansea University
Prifysgol Abertawe

- Academic collaborators (CIs)
  - Irena Spasic, Computer Scientist
  - Jeremy Evas, Welsh Language Expert
  - Paul Rayson, Computational/Corpus Linguist
  - Mark Stonelake, Welsh Language Expert
  - Enlli Thomas, Education and Welsh Language

Lancaster University

Swansea University
Prifysgol Abertawe

PRIFYSGOL
BANGOR
UNIVERSITY

3. CorCenCC: team

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- RAs
  - **Gareth Watkins** – PhD in Translation Tools and Technologies in the Welsh Language Context
  - **Steven Neale** – PhD in Computing, expertise in Natural Language Processing, creative technologies
  - **Jennifer Needs** – PhD in Welsh language teaching (development of online learning materials)
  - **Mair Rees** – PhD in Welsh Literature, expertise in innovative art therapy, creative editor, Gomer Press
  - **Scott Piao** – PhD in Corpus Linguistics, expertise in Corpus Linguistics, Natural Language Processing (NLP) and Text Mining



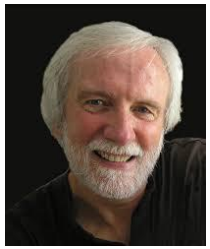


- PhD students: 1 @Cardiff, 1@Swansea (to be recruited)

Tom Cobb, St Louis USA

Kevin Scannell, Missouri USA

Kevin Donnelly Bangor



Michael McCarthy University of Nottingham

Margaret Deuchar University of Cambridge

Laurence Anthony Waseda University, Japan

Consultants

Partners /Stakeholders

Emyr Davies, **CBAC-WJEC**

Gareth Morlais, **Welsh Goverr**

Aran Jones, **SaySomethingI**

Andrew Hawke, **Welsh National**

Dictionary

Owain Roberts, **National Library of**

Meri Huws, **Welsh Language**

sioner

Mair Parry-Jones, **Translation Unit, National Assembly for Wales**

3. CorCenCC: innovation

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- **First large-scale, freely available corpus of Welsh language**
- First **semantic tagger** of Welsh, novel **part-of-speech tagset**
- First Welsh corpus to test **community crowdsourcing** (via an app) for data collection
- **User-defined** corpus, integrating traditional corpus tools with bespoke applications (e.g. the pedagogic toolkit)
- Future-proofed: in-built **sustainability** via an online repository system
- **Building capacity in applied linguistics research** in Wales
- **Model** of corpus construction for under-resourced languages

3. CorCenCC: work packages

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

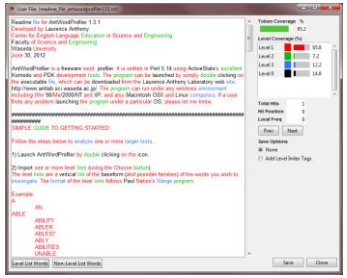
Key work packages:

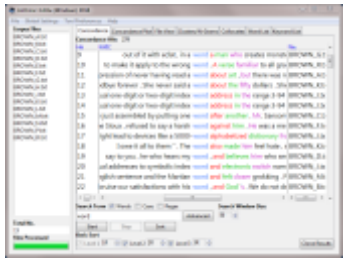
- 1: Collect, transcribe and anonymise the data
- 2: Develop the part-of-speech tag-set/tagger
- 3: Construct semantic annotation software and tagset
- 4: Scope/construct the online pedagogic toolkit www.lextutor.ca/

3. CorCenCC: innovation

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- **CorCenCC** will include a teaching and learning framework
  - Vocabulary profiling tools similar to...
    - Compleat Lexical Tutor (Cobb, 2016)
    - AntWordProfiler (Anthony, 2014)
  - Vocabulary frequency and keyword comparison tools
  - Language 'awareness raising' tools
    - Key-Word-In-Context (KWIC) searches
    - collocations and multi-word unit (MWU) analysis
  - Vocabulary level and size tests





3. CorCenCC: work packages

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Key work packages:

- 1: Collect, transcribe and anonymise the data
- 2: Develop the part-of-speech tag-set/tagger
- 3: Construct semantic annotation software and tagset
- 4: Scope/construct the online pedagogic toolkit www.lextutor.ca/
- 5: Construct infrastructure to host CorCenCC and build the corpus

3. CorCenCC: applications

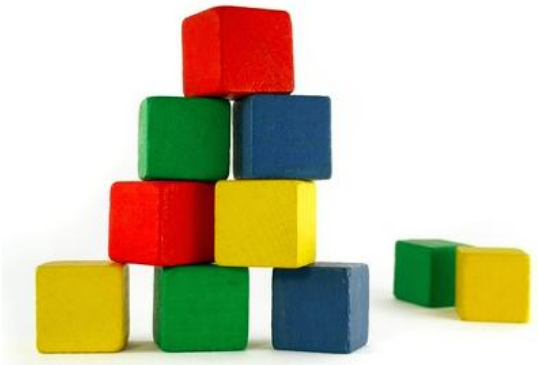
Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- (Some) Potential applications:
  - Pedagogical users
    - Welsh medium education
    - English medium education
    - Welsh for adults
  - Publishers of books and periodicals
  - Print and broadcast media
  - The translation industry
  - Lexicographers

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

A. Planning and piloting
B. Sampling
C. (Meta)data extraction and anonymisation
D. Classification/tagging
E. Visualisation and analysis: constructing and corpus infrastructure



4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

A. Planning and piloting

- Can be a challenge as a 'population without limits, and a corpus is necessary finite at any one point' (Sinclair, 2008: 30) so it is impossible to create a 'complete picture' of discourse in corpora (Thompson, 2005, also see Ochs, 1979; Kendon, 1982: 478-9; Cameron, 2001: 71).

- This is true regardless of whether the corpus is of a specialist or of a more 'general' nature.

- Think about: users and developers, type, purpose, size, representativeness and balance.

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

A. CorCenCC pilot e-language corpus project (2013): why?

- Provided the proof of concept for the wider CorCenCC project
- Ethical considerations/permissions - prompt and positive responses supported our vision of corpus creation as a community enterprise in the Welsh context
- Good opportunity to demonstrate ways in which corpus data can inform prescriptive/descriptive debates: many instances of code-switching and lexical borrowing

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

A. CorCenCC pilot e-language corpus (2013): how?

- Contacted prolific Welsh language tweeters and bloggers via email and sought permission to use material to ensure sites were likely to be read by a critical mass of Welsh speakers, so as to be representative of 'typical' online Welsh language.
- [NB CorCenCC does not include tweets – usage rights preclude publication)
- Used API to extract data
- Indexed > database > anonymisation
- Scrutinised data for specific features

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

B. Sampling: balance and representativeness

- **Lessons learned from the CorCenCC pilot:**
- The actual number gained was determined by the following factors, the majority of which were beyond the control of the corpus developers:
  - The targeted number of words to collect for each type;
  - The rate at which a user publishes content;
  - The size of contributions;
  - The time over which they are collected.

Mode	Contributors	Individual entries	Word count
Blogs	40	810	257,658
Tweets	32	17,460	271,650
	72	**18,270**	**529,308**

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

B. Sampling: balance and representativeness

- *CorCenCC will be a **general** corpus so will include data sampled from a range of different speakers (of different ages and occupations), across a range of different discourse contexts, and geographical locations of Wales. This will allow users to make generalised observations about language use (i.e. not restricted to a specific discourse context or domain).*

- *It will be **balanced** and **representative**.*

- **Q:** What questions can we actually ask about Welsh using CorCenCC?

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

B. Sampling: balance and representativeness

- Is **balance** and **representativeness** actually ever possible? Probably not.

- The key thing is not about representativeness and balance but about the predictive power of a model. Anyone can create a model – it is not the model that is important but what it can do and the predictive power it has.

- Most CL is purely descriptive and about the past - description needs to be extended to think about the future.

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

B. Sampling: challenges…e-language and beyond

- **Demographics – e.g. age**
  - Young people: very important age group (over 27% of speakers are under 15 – 2011 census), but ethics of data collection?
- **Location**
  - Areas where Welsh speakers are in a very small minority (e.g. less than 1% of the population): sparseness of data?
- **Text genres**
  - Some genres used by the BNC, for example, not relevant for Welsh
  - E-language: enough blogs/websites to get adequate coverage of all genres?

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

B. Sampling: CorCenCC 'proper' – blogs

Thema/Pwnc / *Theme/Topic*	**%**	**Geiriau / *Words***
Blog	**30%**	**600,000**
A: Newyddion, Y Cyfryngau a Materion Cyfoes / *News, Media and Current Affairs*, Gwleidyddiaeth / *Politics*, Busnes a Chyllid / *Business and Finance*, Y Tywydd a'r Amgylchedd / *Weather and the Environment*, Siopa Ar-lein / *Online Shopping*	5%	100,000
B: Crefydd / *Religion*, Yr Iaith / *Language*, Diwylliant, Llenyddiaeth a'r Celfyddydau / *Culture, Literature and the Arts*, Addysgu, Academia ac Addysg / *Teaching, Academia and Education*	5%	100,000
C: Technoleg, Cyfrifiaduron a Chwarae Gemau Cyfrifiadurol / *Technology, Computers and Gaming*, Ffasiwn a Harddwch / *Fashion and Beauty*, Hobïau a Difyrrwch / *Hobbies and Pastimes*, Teithio / *Travel*, Coginio / *Cookery*	5%	100,000
D: Cerddoriaeth / *Music*, Chwaraeon / *Sport*, Perfformiadau byw a Digwyddiadau / *Gigs and Events*	5%	100,000
E: Hynt a Helynt Pobl Enwog / *Celebrity news and gossip*, Teledu a Ffilm / *TV and Film*, Hiwmor / *Humour*	5%	100,000
F: Bod yn Rhiant a Bywyd Teuluol / *Parenting and Family Life*, Iechyd a Lles / *Health and Wellbeing*, Bywyd Personol a Phob Dydd / *Personal and Daily Life*	5%	100,000

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

B. Sampling: CorCenCC 'proper' – websites

Gwefan / *Website*	30%	600,000
A: Newyddion, Y Cyfryngau a Materion Cyfoes / *News, Media and Current Affairs*, Gwleidyddiaeth / *Politics*, Busnes a Chyllid / *Business and Finance*, Y Tywydd a'r Amgylchedd / *Weather and the Environment*, Siopa Ar-lein / *Online Shopping*	5%	100,000
B: Crefydd / *Religion*, Yr Iaith / *Language*, Diwylliant, Llenyddiaeth a'r Celfyddydau / *Culture, Literature and the Arts*, Addysgu, Academia ac Addysg / *Teaching, Academia and Education*	5%	100,000
C: Technoleg, Cyfrifiaduron a Chwarae Gemau Cyfrifiadurol / *Technology, Computers and Gaming*, Ffasiwn a Harddwch / *Fashion and Beauty*, Hobïau a Difyrrwch / *Hobbies and Pastimes*, Teithio / *Travel*, Coginio / *Cookery*	5%	100,000
D: Cerddoriaeth / *Music*, Chwaraeon / *Sport*, Perfformiadau byw a Digwyddiadau / *Gigs and Events*	5%	100,000
E: Hynt a Helynt Pobl Enwog / *Celebrity news and gossip*, Teledu a Ffilm / *TV and Film*, Hiwmor / *Humour*	5%	100,000
F: Bod yn Rhiant a Bywyd Teuluol / *Parenting and Family Life*, Iechyd a Lles / *Health and Wellbeing*, Bywyd Personol a Phob Dydd / *Personal and Daily Life*	5%	100,000

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

B. Sampling: CorCenCC 'proper' – email and SMS

Ebost / *Email*		**20%**	**400,000**
Proffesiynol *Professional*	e.e. ebost i gadarnhau amser cyfarfod *e.g. an email to confirm a meeting*	13.4%	268,000
Personol *Personal*	e.e. ebost sy'n rhannu newyddion da *e.g. an email to share good news*	6.6%	132,000

Negeseuon Testun Electronig Byr / *Short Electronic Text Messages*		**20%**	**400,000**
Proffesiynol *Professional*	e.e. Neges sydd wedi ei hanfon gan ysgol sy'n darparu gwybodaeth ynghylch noswaith rhieni / *e.g. a message sent by a school providing details of a parents' evening*	6.6%	132,000
Personol *Personal*	e.e. neges ynghylch cwrdd a ffrind am goffi *e.g. a message regarding meeting a friend for coffee*	13.4%	268,000
		100%	**2,000,000**

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. (Meta)data extraction and anonymisation

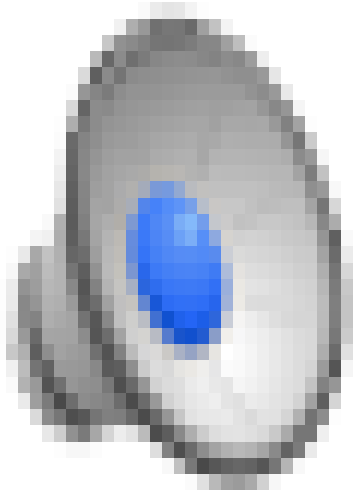
- Semi-automated techniques to be utilised?

- Possible techniques = automated extraction using APIs
  - http://bootcat.sslmit.unibo.it/
  - http://www.tweepy.org/ - Python library for accessing the Twitter API.
  - https://www.facebook.com/birdbodycorpus/posts/5063944?hc_location=ufi

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. (Meta)data extraction and anonymisation



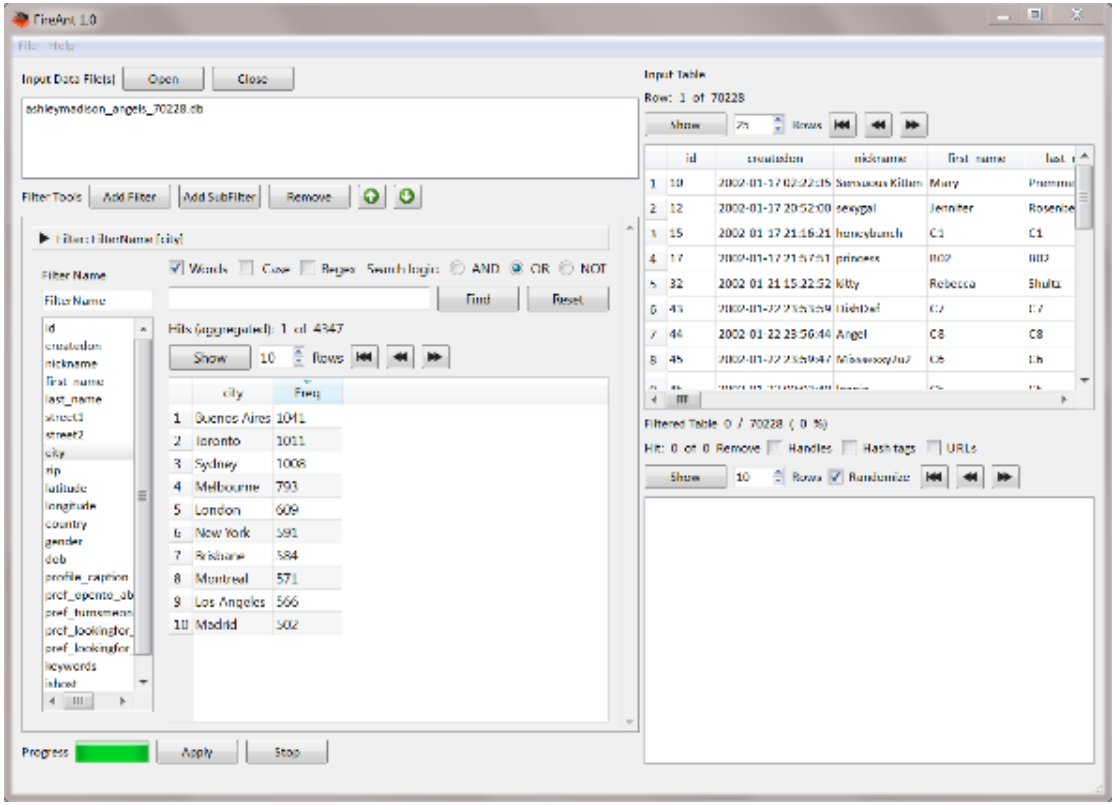
www.cs.cf.ac.uk/cosmos/

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. (Meta)data extraction and anonymisation

- Fireant - http://www.laurenceanthony.net/software/fireant/ - "[F]ilter, [I]dentify, [R]eport & [E]xport [An]alysis [T]oolkit"



4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. (Meta)data extraction and anonymisation

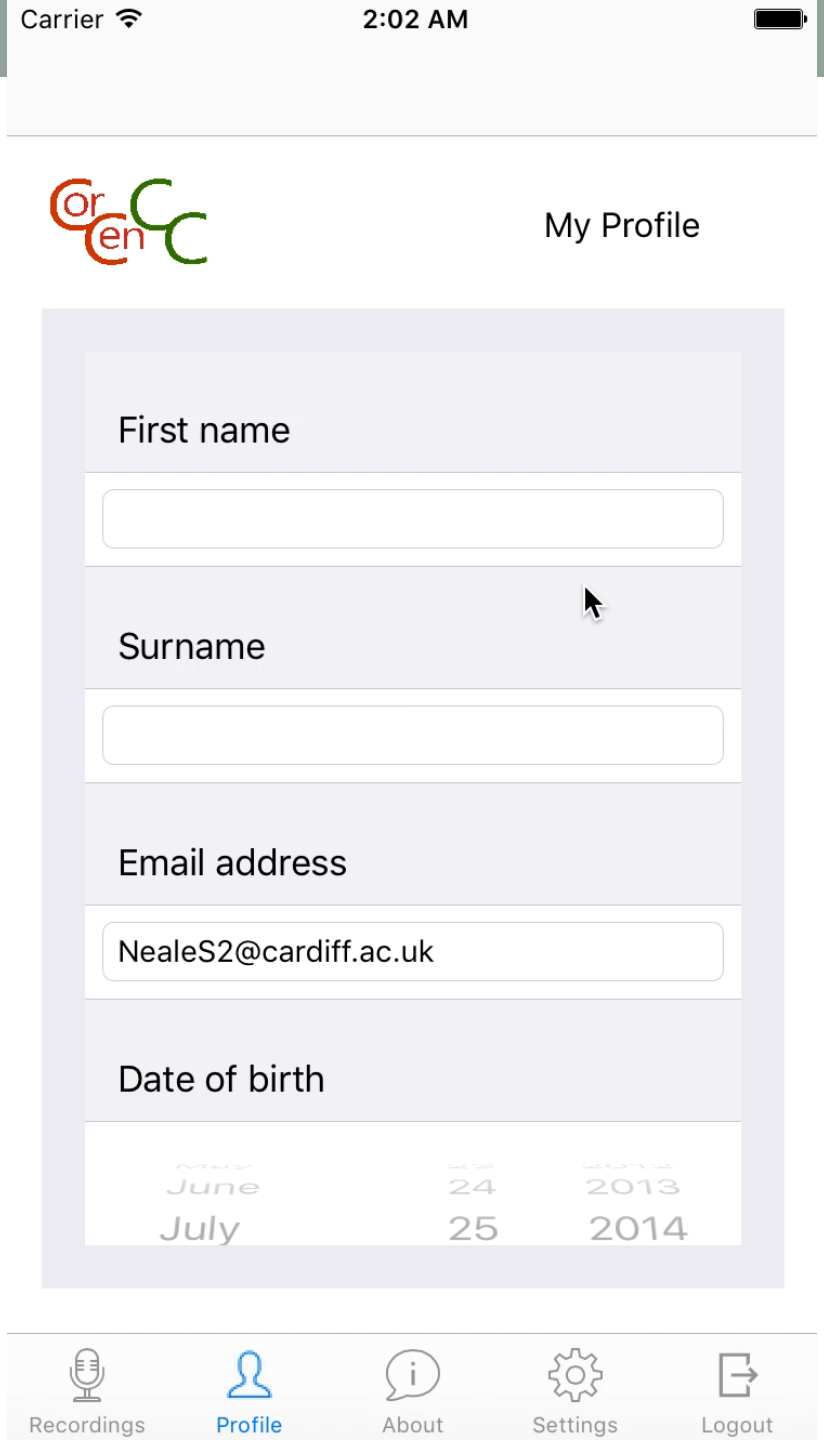
Crowdsourcing other forms of data collection:

- Crowdsourcing – an 'online, distributed problem-solving and production model' (Brabham, 2008: 75) involving 'internet-based collaborative activity, such as co-creation and user innovation' (Estellés-Arolas, 2012: 189).

- The out**sourcing** of tasks and activities to groups and networks of people (**crowd**).

- The use of crowdsourcing will facilitate the engagement of future users of the corpus from the very start of its development (a user-driven corpus design).

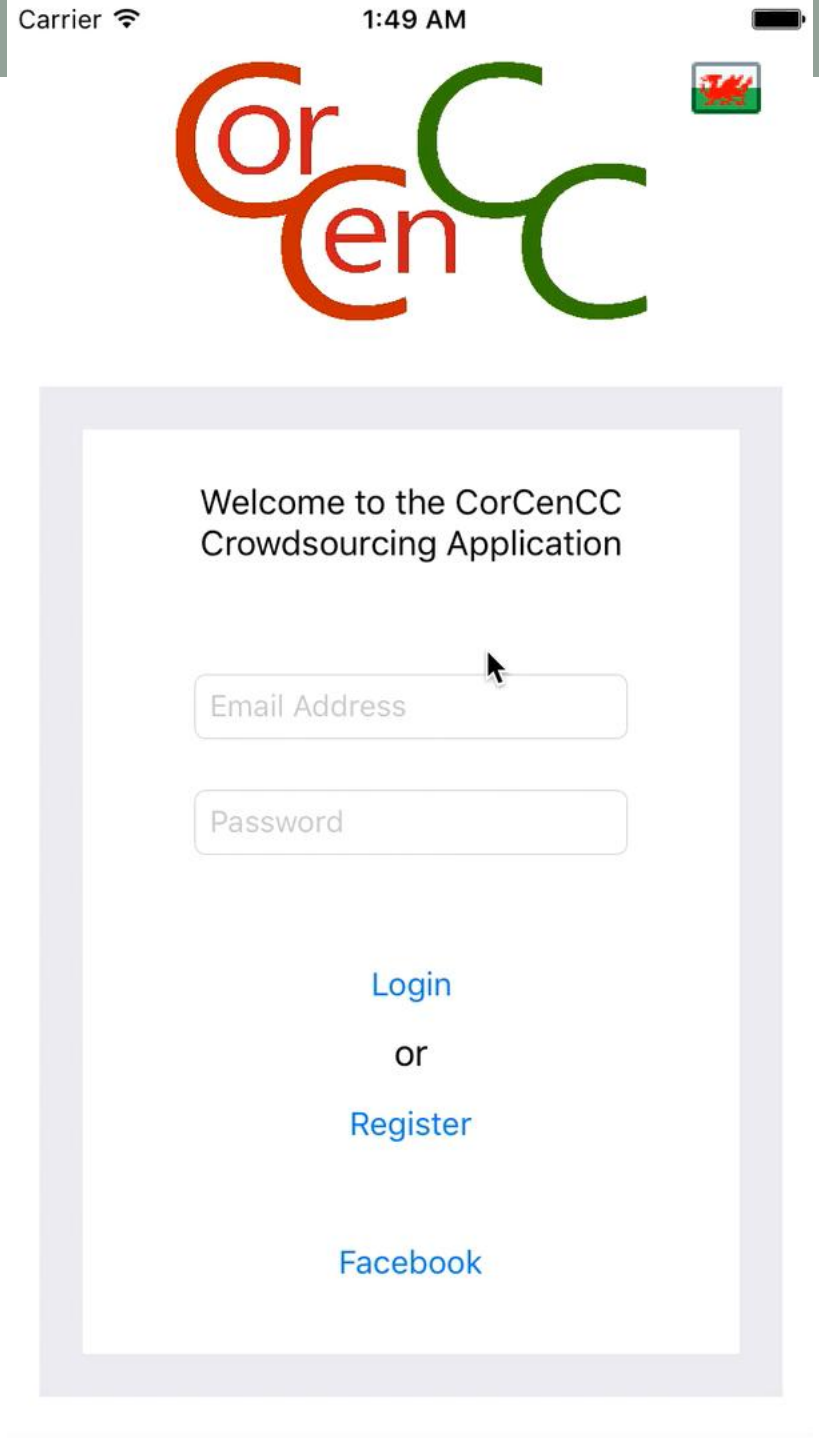
Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh



- Risks
- Public buy-in
- Signal problems
- Accessibility

Based on a pilot app – many thanks to Newcastle University

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh



- Risks
- Public buy-in
- Signal problems
- Accessibility

Based on a pilot app – many thanks to Newcastle University

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. (Meta)data extraction and anonymisation

- Including a complete set of metadata for all e-language types may be difficult, if not impossible.

- While contributors of short electronic text messages and email messages can be asked to provide data in respect of age and gender, for instance, the same information cannot necessarily be ascertained for blogs and websites. It is true that, as Schler et al. (2006: 1) note, 'many […] blogs include formatted demographic information provided by the authors'.

- COSMOS 'predicted' genders…

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. Anonymisation

- E.g. BAAL 'Recommendations on Good Practice in Applied Linguistics' (page 5)
  - *'In some cases, such as participatory or collaborative research with professionals and some forms of internet research, anonymity may be impossible or or unfavourable, as where an internet site's regulations state that data should not be altered, or where an author, or joint practitioner/researcher, wishes to be acknowledged. In such cases, specific regulatory frameworks governing research sites, and/or the autonomy of individual informants, must be negotiated.'*

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. Anonymisation

CorCenCC Anonymization Conventions

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

CONTENTS

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

C. Anonymisation

Section	Aspect of Anonymization	Convention
1.	SENSITIVE DATA	**All instances of sensitive data should be logged, noting that the data is believed to be sensitive (see ADDITIONAL)** Special care needs to be taken when data is sensitive. Data can be considered sensitive if it: • is a potential libel (Merriam Webster define libel as follows: 'the act of publishing a false statement that causes people to have a bad opinion of someone') • is a potential slander (Merriam Webster define slander as follows: 'to make a false spoken statement that causes people to have a bad opinion of someone') • could potentially cause someone embarrassment (e.g. revealing personal information) • could potentially cause someone financial loss (e.g. divulging bank details) • could potentially put someone in a compromising position (e.g. discussion of misconduct at work or illegal activities) • has potential to cause someone harm in any other way (e.g. divulging passwords)

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

D. Classification/tagging

Processing uploaded data:

- Pre-processing:
  - Convert; clean; strip/extract; anonymization [1]; editing
- Natural Language Processing (NLP) steps:
  - Part-of-speech (POS) tagging; semantic category tagging

	the	cat	sat	on	the	mat
POS	DT	NN	VBD	RP	DT	NN
Sem		L1				H5

- Post-processing:
  - Anonymization [2]

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

D. Classification/tagging

- Bespoke POS Tagset for Welsh – coming soon

Rhan Ymadrodd/POS	Tag CorCenCC Tag	Disgrifiad_cy	Description_en
Enw	Egu	Enw gwrywaidd unigol	Noun, common masc. Singular
	Ebu	Enw benywaidd unigol	Noun, common fem. Singular
	Egll	Enw gwrywaidd lluosog	Noun, common masc. Plural
	Ebll	Enw benywaidd lluosog	Noun, common fem. Plural
	Egbu	Enw gwrywaidd/benywaidd unigol	Noun, common masc./fem. Singular
	Egbll	Enw gwrywaidd/benywaidd lluosog	Noun, common masc./fem. Plural
	Epg	Enw priod gwrywaidd	Noun, proper, masc.
	Epb	Enw priod benywaidd	Noun, proper, fem.
Y Fannod Benodol	YFB	Y Fannod Benodol	Article, definite
	Arsym	Arddodiad syml	Preposition, uninflected
	Ar1u	Arddodiad rhediadol, pers. 1af. unigol	Inflected preposition, 1st pers. Singular
	Ar2u	Arddodiad rhediadol, 2il pers. unigol	Inflected preposition, 2nd pers. Singular

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

D. Classification/tagging

- Semantic Category Tagset for Welsh – available now

K ENTERTAINMENT, SPORTS & GAMES	
K1	Entertainment generally
K2	Music and related activities
K3	Recorded sound etc.
K4	Drama, the theatre & show business
K5	Sports and games generally
K5.1	Sports
K5.2	Games
K6	Children's games and toys
L LIFE & LIVING THINGS	
L1	Life and living things
L2	Living creatures generally
L3	Plants
M MOVEMENT, LOCATION, TRAVEL & TRANSPORT	
M1	Moving, coming and going
M2	Putting, taking, pulling, pushing, transporting &c.
M3	Movement/transportation: land
M4	Movement/transportation: water
M5	Movement/transportation: air
M6	Location and direction
M7	Places
M8	Remaining/stationary

K ADLONIANT, CHWARAEON A GEMAU	
K1	Adloniant yn gyffredinol
K2	Cerddoriaeth a gweithgareddau cysylltiedig
K3	Sŵn wedi'i recordio ac ati
K4	Drama, theatr a byd adloniant
K5	Chwaraeon a gemau yn gyffredinol
K5.1	Chwaraeon
K5.2	Gemau
K6	Gemau a theganau plant
L BYWYD A PHETHAU BYW	
L1	Bywyd a phethau byw
L2	Creaduriaid byw yn gyffredinol
L3	Planhigion
M SYMUD, LLEOLIAD, TEITHIO A CHLUDIANT	
M1	Symud, dod a mynd
M2	Gosod, cymryd, tynnu, gwthio, cludo ac ati
M3	Cerbydau a chludiant ar dir
M4	Morgludiant, nofio ac ati
M5	Awyrennau a hedfan
M6	Lleoliad a chyfeiriad
M7	Lleoedd
M8	Aros/aros yn yr unfan

- Iterative developments to this tagset using crowdsourcing methods.

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

E. Visualisation and analysis: constructing corpus infrastructure

- **Back-end (repository system):** design and construction of an online system which allows for the introduction of new data to the corpus over time, with the maintenance of the corpus being supported by its own users, making contributions to the corpus a social venture.

- **Front-end (corpus infrastructure):** includes KWIC (Key Word in Context) concordancers and collocation tools, search and sort tools, word frequency lists, key word analysers and statistical testing facilities. Users will also be able to search for and replay audio files and visualise data.

4. Corpus design and construction

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

http://wordwanderer.org

5. Ethical considerations

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- Baker and McEnery note (2015: 246-7) ‘as a new form of language use, ethical practices when carrying out research in social media are continually developing and there is no current common consensus around ‘best practice’’. This on-going change can prove to be particularly problematic when planning and developing datasets for analysis.

- ‘Ethics’ at multiple levels including: National; Institutional; Funding-councils; Discipline-specific; personal..

Name	Size	Reference	Notes
Edinburgh Twitter Corpus	97 million Twitter posts, over 2 million words	Petrović et al., 2010	Taken down.
HERMES general English Language corpus	7 million Twitter posts, 100 million words	Zappavigna, 2012	Not publically available.
Rovereto Twitter N-Gram Corpus	75 million posts from 11 million users from 2010-2011 (taken from a larger dataset of 240 million tweets)	Herdağdelen and Baroni, 2011	Metadata available but not full tweet texts.
Stanford Twitter Corpus	467 million posts from 20 million users over 7 months in 2009	Yang and Leskovec, 2011	Taken down.
TREC Tweets 2011	16 million tweets collected from 23rd January to 8th February 2011	Horn et al., 2011	Identifiers, not data, released.
Twitter_Smallcorp	2 million words	Puschmann, 2009	Not publically available.

5. Ethical considerations

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

- E.g. Twitter - while it is not possible to distribute data away from the Twitter site, it is permissible to distribute metadata from tweets, including the time and date that they were collected, and the Twitter handle (i.e. username) used by the individual Tweeter. These identifiers can then be used by other researchers to collect and reconstitute the dataset for themselves at a later date. This is prone to high levels of decay.

- The fluidity of 'terms of service'
  - https://www.youtube.com/watch?feature=player_embedded&v=Aifb49urxKM
  - https://tosdr.org/

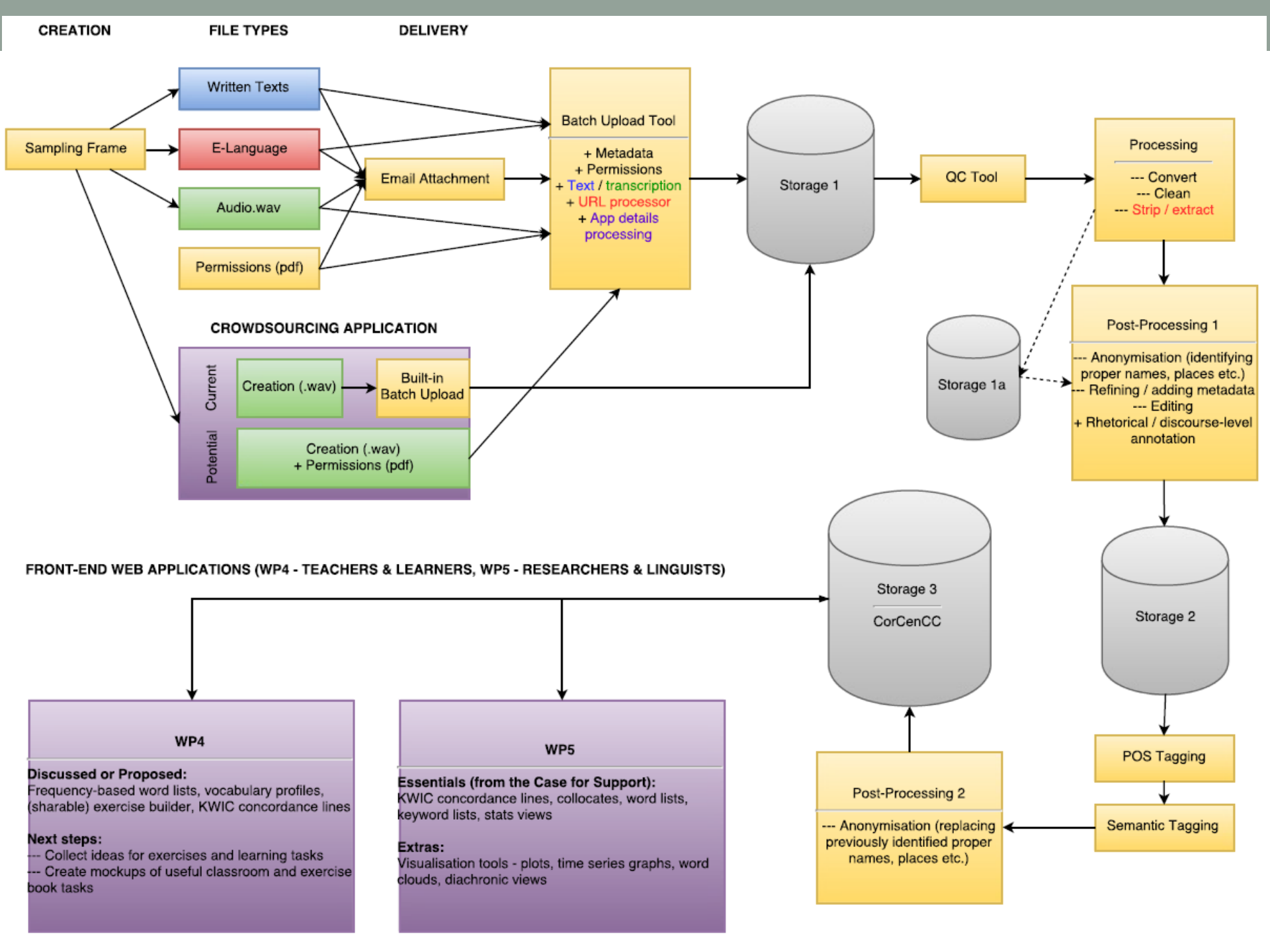
5. Ethical considerations

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Permission workflow for CorCenCC

Stage	Documents for contributors	Specific document to use[1]	Language data					Surveys (online and hard copies)
			Spoken	Written	E-Language			
					Websites/Blogs	SMS/E-mail		
Stage 1	**Provision of generic project information (initial contact)**	Leaflet	**X**	X	X	X		**X**
		E-mail[2]	**X**	X	X	X		**X**
		Website	X	**X**	X	X		**X**
		App	X	**X**	**X**	**X**		
Stage 2	**Permissions (contributors to complete)**	Hard copy permission form[3] – (full signature required)	**X**	**X**	**X**	**X**		**X**
		Online registration process – box to be ticked to signal agreement	**X**	X	X	**X**		X
		App permission process – box to be ticked to signal agreement	**X**					
Stage 3	**Contributor information (contributors to complete)**	Online registration process	X	**X**	**X**	**X**		X
		App registration process	**X**					
		Contributor information form[4]	**X**	**X**	**Not required – sampling by theme**	**X**		**X (embedded in survey)**

*boldface indicates the 'typical'/most common way in which these processes are carried out.


CREATION
FILE TYPES
DELIVERY
Written Texts
Sampling Frame
E-Language
Audio.wav
Permissions (pdf)
Email Attachment
Batch Upload Tool
+ Metadata
+ Permissions
+ Text / transcription
+ URL processor
+ App details processing
Storage 1
QC Tool
Processing
--- Convert
--- Clean
--- Strip / extract
CROWDSOURCING APPLICATION
Current
Creation (.wav)
Built-in Batch Upload
Potential
Creation (.wav) + Permissions (pdf)
Storage 1a
Post-Processing 1
--- Anonymisation (identifying proper names, places etc.)
--- Refining / adding metadata
--- Editing
+ Rhetorical / discourse-level annotation
Storage 3
CorCenCC
Storage 2
FRONT-END WEB APPLICATIONS (WP4 - TEACHERS & LEARNERS, WP5 - RESEARCHERS & LINGUISTS)
WP4
Discussed or Proposed:
Frequency-based word lists, vocabulary profiles, (sharable) exercise builder, KWIC concordance lines
Next steps:
--- Collect ideas for exercises and learning tasks
--- Create mockups of useful classroom and exercise book tasks
WP5
Essentials (from the Case for Support):
KWIC concordance lines, collocates, word lists, keyword lists, stats views
Extras:
Visualisation tools - plots, time series graphs, word clouds, diachronic views
Post-Processing 2
--- Anonymisation (replacing previously identified proper names, places etc.)
POS Tagging
Semantic Tagging

6. Reflections/future directions

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

http://sites.cardiff.ac.uk/corcencc/

www.corcencc.org

www.facebook.com/CorCenCC/

@CorCenCC

For more information, visit one of the sites above, or email the project team:

CorCenCC@Cardiff.ac.uk