



# **The Linguistic Data Consortium: Developing and Sharing Resources for Indigenous Languages**

Christopher Cieri, Denise DiPersio

University of Pennsylvania, Linguistic Data Consortium

{ccieri, dipersio} AT ldc.upenn.edu

- ◆ LDC: Founding and Mission
- ◆ Sharing Data in the World's Languages
- ◆ Research Collaborations in Indigenous Languages
- ◆ Novel Data Collection Methods for Indigenous Languages

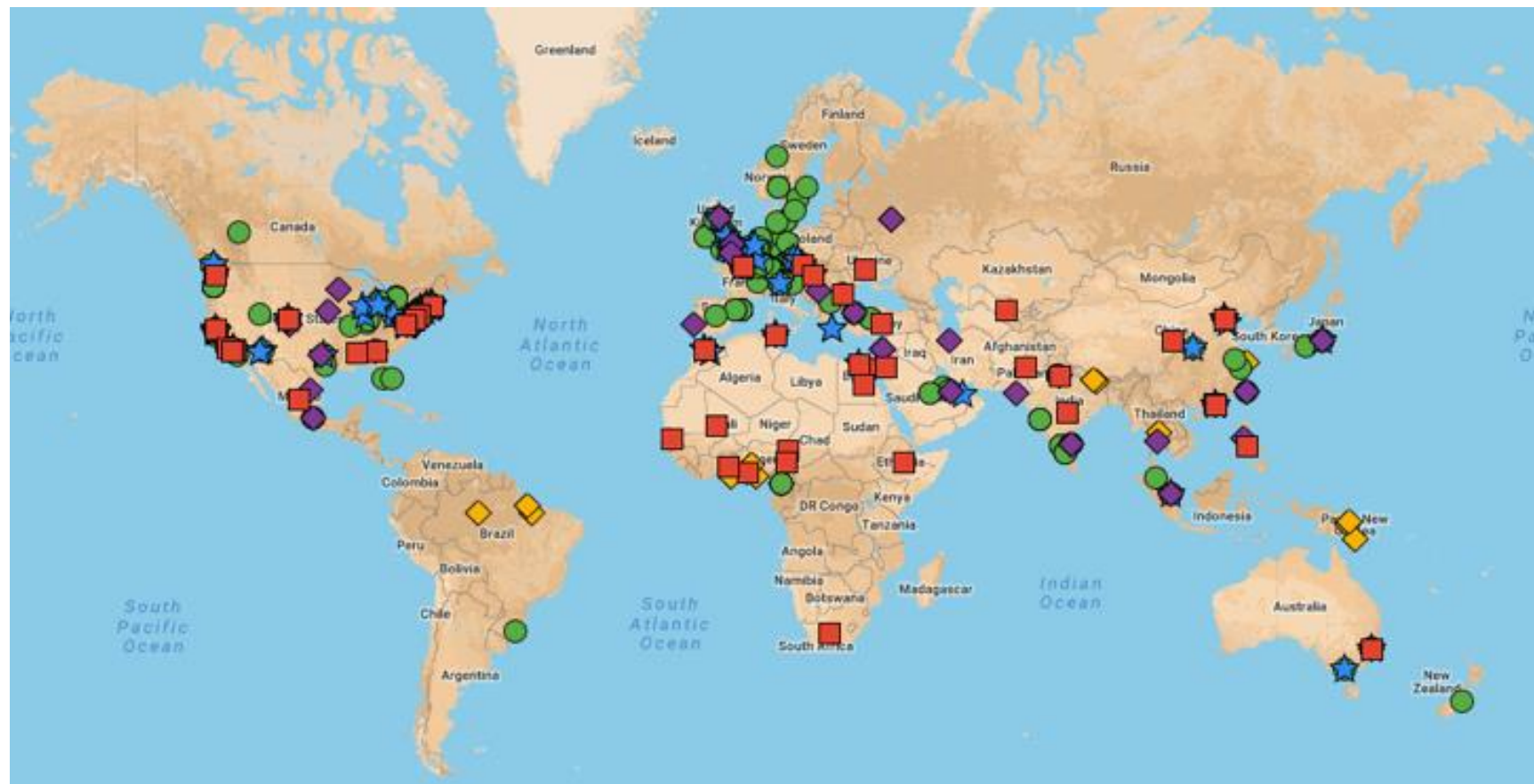
- ◆ A mutual aid society with the mission to develop and distribute language resources to the global community
  - Academia, government, industry
  - Researchers contribute data sets: visibility, community recognition, uptake
  - Members/data licensees contribute fees: ongoing rights to a variety of resources, high ROI
  - Sponsors contribute funding: resource creation, infrastructure, innovation, cost sharing, resource dissemination to the community
- ◆ LDC's online Catalog launched in 1993
  - Over 190,000 copies of 820+ resources in more than 90 languages distributed to roughly 6000 distinct organizations in over 100 countries
  - 3-4 new data sets released monthly
  - Distributed under a variety of licensing arrangements: for use in language-related research, education and technology development

- ◆ The LDC Catalog is a permanent language resource archive
  - Seeded by data contributions of significant corpora, augmented by data sets developed by LDC in funded projects along with contributions from the global research community
- ◆ The Catalog is a CoreTrustSeal trustworthy repository
  - Meets high standards for data access, rights management, curation, storage, security
- ◆ Metadata and catalog descriptions follow established standards and best practices for digital repositories
- ◆ New publications announced in LDC's monthly newsletter sent to over 20,000 recipients
- ◆ Research impact: more than 10,000 papers cite LDC data
- ◆ LDC has the expertise and infrastructure to ensure that data is preserved and accessible, with appropriate protections to indigenous language communities, students, scholars, researchers and developers

- ◆ text, image, audio, video, multimedia data from news sources, journals, financial, government, biomedical docs
- ◆ internet sources including newsgroups, (micro)blogs and discussion fora
- ◆ text interactions via email, chat and SMS
- ◆ machine printed, handwritten and hybrid document/document images
- ◆ audiovisual data from broadcast news and conversation, podcasts, conversational telephone speech, lectures, interviews, meetings, field interviews, read and prompted speech, task oriented speech, clinical evaluations, games & role play, speech in noise, web video and even animal vocalizations (elephant, zebra finch)
- ◆ selfies, self-produced images, videos and multimedia data
- ◆ digitized analog media including interviews in a variety of tape formats

- ◆ data scouting, data triage and smart data selection
- ◆ alignment of paired audio, auditing of bandwidth, signal quality, language, dialect, program, speaker
- ◆ quick, quick-rich and careful transcription, audio segmentation and audio-text alignment at story, turn, sentence, word level
- ◆ orthographic, spelling and phonetic script normalization and transliteration
- ◆ tagging of phonetic, dialect, sociolinguistic and supralexic features
- ◆ text localization, document zoning, handwritten and machine print image transcription, OCR post-editing, tagging of reading order
- ◆ word and sentence segmentation, segmentation and tagging of morphology, part-of-speech and gloss, NP chunking, Treebanking, PropBanking, AMR SemBanking
- ◆ sense disambiguation, fine and coarse-grained topic relevance annotation
- ◆ text and multimedia novelty, entailment, hypothesis generation and inference annotation
- ◆ annotation of committed belief, sentiment, emotion, disfluency, hedging, discourse features
- ◆ detection, classification, coreference, knowledge base population of entities, events, acts, states, situations, relations, time, location, other attributes in text, image, audio, video, multimedia data
- ◆ single and multi-document summarization of various lengths from titles to 200 words
- ◆ query generation and question answering
- ◆ (multiple) translation, edit distance, post-editing, paraphrasing, HyTER, translation quality control
- ◆ alignment of translated text at document, sentence, phrase, word and morpheme levels
- ◆ describing the physics of gesture via joint angles and rotations
- ◆ identification, classification and tracking entities and events in video
- ◆ assessment of IR, MT, KBP, QA, hypothesis, sentiment and other system output





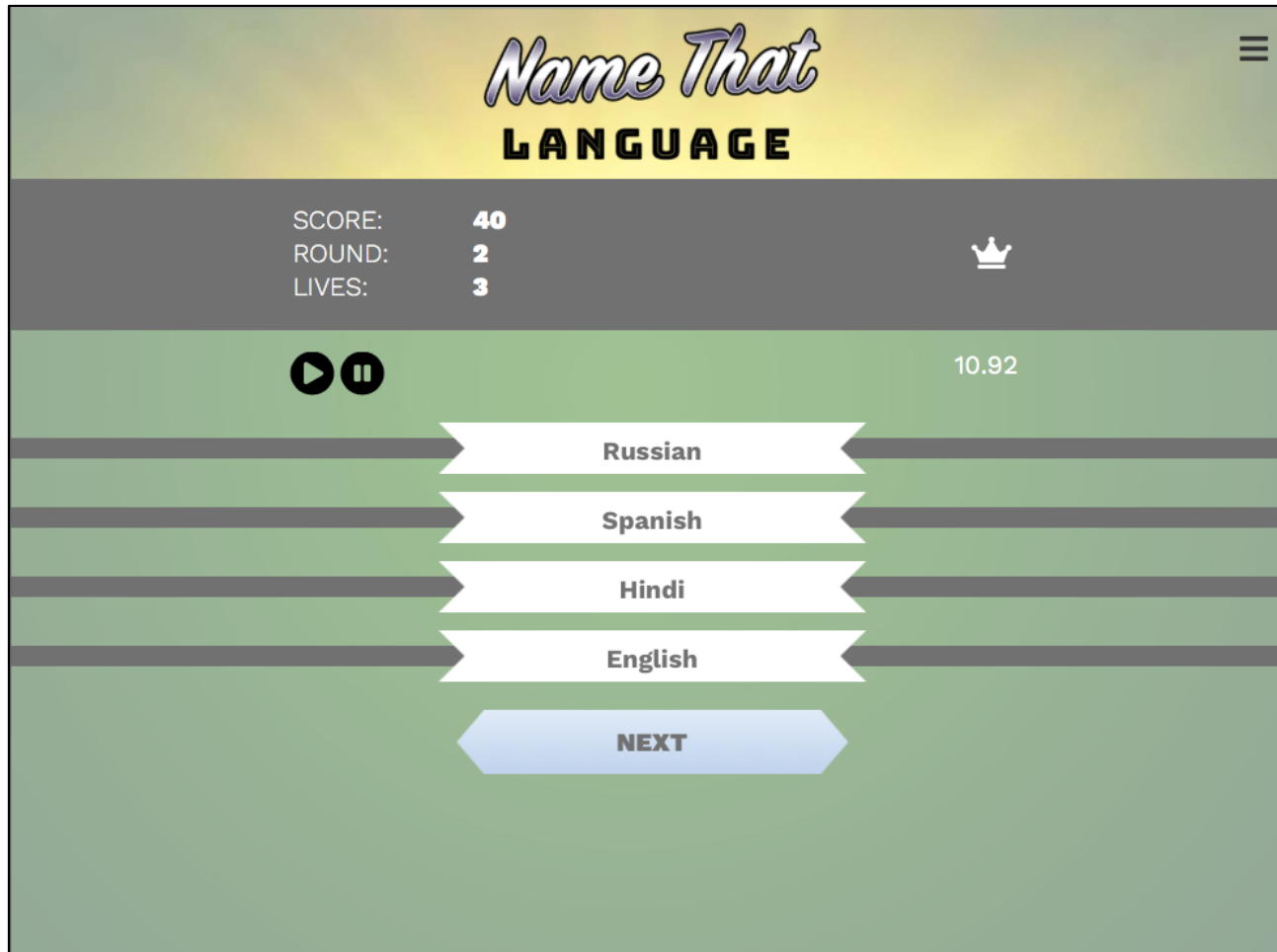
LDC Global Network of select data sources including: ■ = subcontractors and vendors, ● = corpus authors, ◆ = media providers, ◆ = LDC staff collections, ★ = research collaborators. Many markers represent multiple collaborators; many markers partially obscured by others.

# Research Collaborations in Indigenous Languages


- ◆ Language documentation support
  - AARDVARC (Automatically Annotated Repository of Digital Audio and Video Resources Community)
  - EMELD (Electronic Metastructure for Endangered Languages Data)
- ◆ Advice and technical assistance for collections of Nahuatl, Mixtec, Tembé and Nhengatu
- ◆ LDC workshops around languages in the Americas
  - Philadelphia 2018: Planning Workshop on Data Archives and Languages of the Americas
    - Experts managing linguistic data archives and resource centers discussing challenges, needs and opportunities for promoting and extending collaboration in the Americas
  - Mexico City 2018: International Workshop on Data Intensive Research on Languages of the Americas
    - Linguists and scientists from Mexico, Brazil, Chile, Argentina, USA
    - Languages discussed include Chuj, Yucateco, Huasteco, Nahuatl, Wixarika, Southern Cone languages, Mexican/American Spanish, Brazilian Portuguese



\*Mobile friendly; better visual appeal






- data type, source
- language selection
- # instances/language
- distractor selection
- # distractors
- game play variants
- scoring
- feedback
- competition



[ABOUT US](#)
[JOIN US](#)
[PROJECTS](#)
[CREATE A PROJECT](#)
[NEWS](#)
[CHAT](#)

[USERTEST](#)
[SIGN OUT](#)

Language Analysis Research Community


A citizen science community for research in language, linguistics and machine learning.

[Learn More](#)


**JOIN US!**  
Help make the world smarter!

[Register](#) [View Projects](#)


FEATURED PROJECTS



TONGUE TWISTERS




ITALIAN: DIALECTS, REGIONAL AND STANDARD



SPEECH BIOMARKERS

[ALL PROJECTS](#)

PARTNERS



Language)Arc

JFLMARA | SIGN OUT

ABOUT US

JOIN US

PROJECTS

CREATE A PROJECT


NEWS

CHAT


Projects

Sort By:


Search




**Tongue Twisters**  
 Record tongue twisters; identify and classify speech errors in others' recordings




**Italian: Dialects, Regional and Standard**  
 Help document the current differences among Standard and Regional Italian and Dialects.




**Speech Biomarkers**  
 Tracking Speech as a Wellness Biomarker



**The Dark Triad Survey**  
 Help us create a general population dataset for personality trait research.



**Understanding Autism Spectrum Disorder**  
 Help us create a general population baseline dataset for autism research.



**Elicitation Corpus Translation**  
 Help us create a multilingual translation corpus.

Portions © 2019 University of Pennsylvania

Funded in part by a grant from the National Science Foundation

ABOUT US

FAQ

TERMS OF SERVICE

CONTACT

f

in



## How Language Varies

### Reading Passages

[ABOUT](#)[ANNOTATE](#)[CHAT](#)[Tutorial](#)[Reference](#)

Please record yourself reading the passage in your normal voice and style.

Please call Stella

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

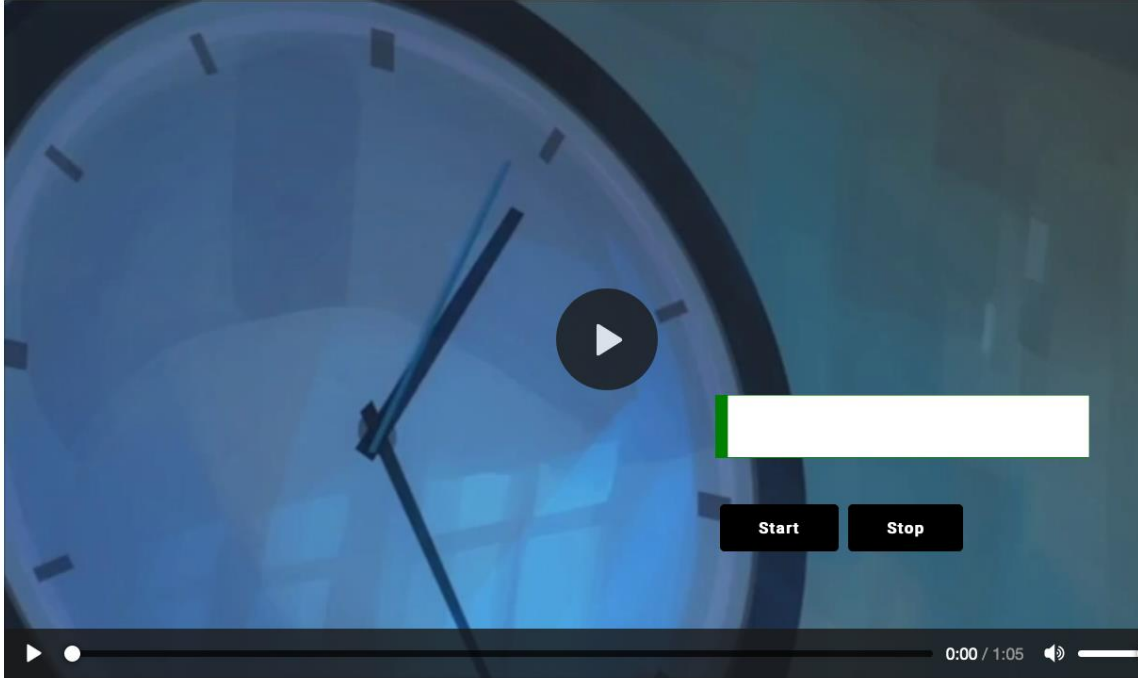
[Start](#)[Stop](#)[Submit](#)[Skip](#)[Report](#)

Once upon a time ...

How Language Varies  
Silent Movie Narration

ABOUTANNOTATECHAT

After reading the tutorial on creating a great recording at least once, please record yourself describing the persons, things, places and activities in the movie as if for someone who can not see it.

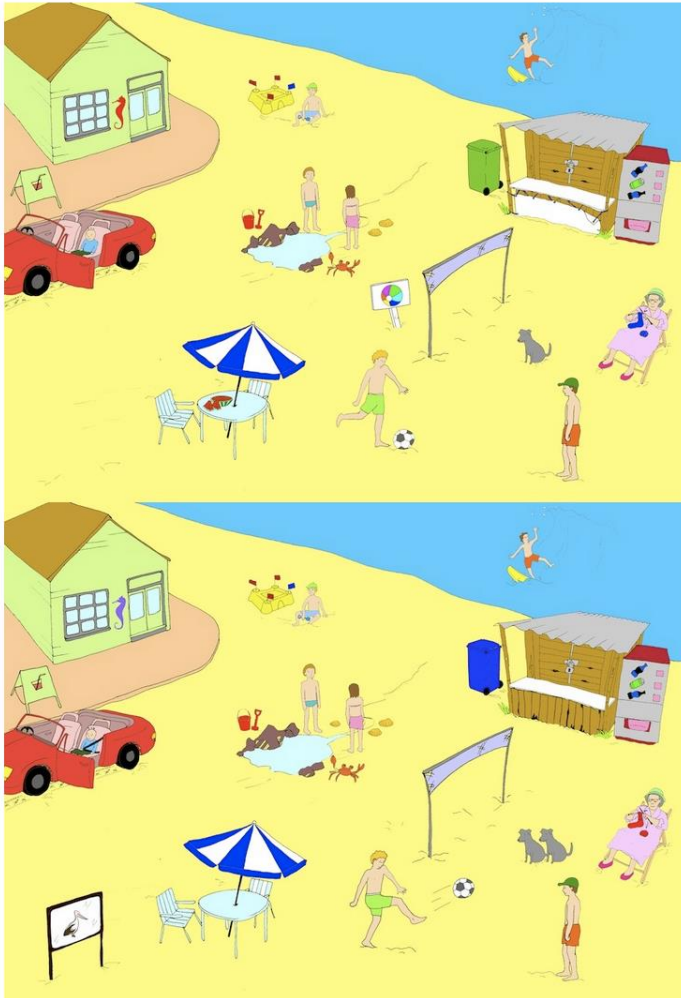


StartStop

0:00 / 1:05

SubmitSkipReport

Record yourself identifying the differences in these two pictures as though for someone who can not see them.



Start

Stop

Submit

Skip

Report



Look at the image and give the local name for what you see in your native variety. If you speak dialect give the dialect name. Otherwise give the name in Italian.



Start

Stop

Optional Written Name:

Submit

Skip

Report



Language)Arc

USERTEST | SIGN OUT

ABOUT US

JOIN US

PROJECTS

CREATE A PROJECT

NEWS

CHAT



## Elicitation Corpus Translation

### Elicitation Task: Primary

ABOUT

ANNOTATE

OUR RESEARCH TEAM

CHAT

Tutorial

Translate the English source sentence into your selected language based on the context

Eastern Farsi (prs)

**Source Sentence:** Michael was greeting Patricia.

**Context:** The speaker asserts this sentence to be true

**Translation:**

Submit

Skip

Report

Language)Arc

USERTEST | SIGN OUT

ABOUT US


JOIN US

PROJECTS

CREATE A PROJECT

NEWS

CHAT



Italian: Dialects, Regional and Standard

Classifying Italian Speech

ABOUT

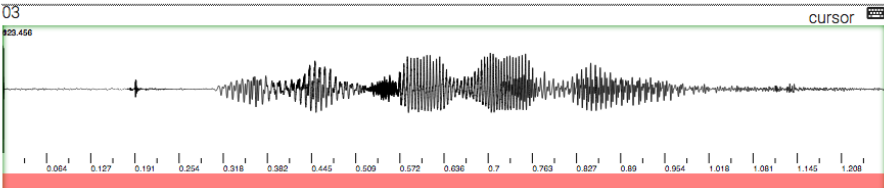
ANNOTATE

OUR RESEARCH TEAM

CHAT

Listen to the audio clip and classify the speech as Standard or Regional Italian or Dialect using the buttons below.

03



1.272

Standard

Regional

Dialect

Skip

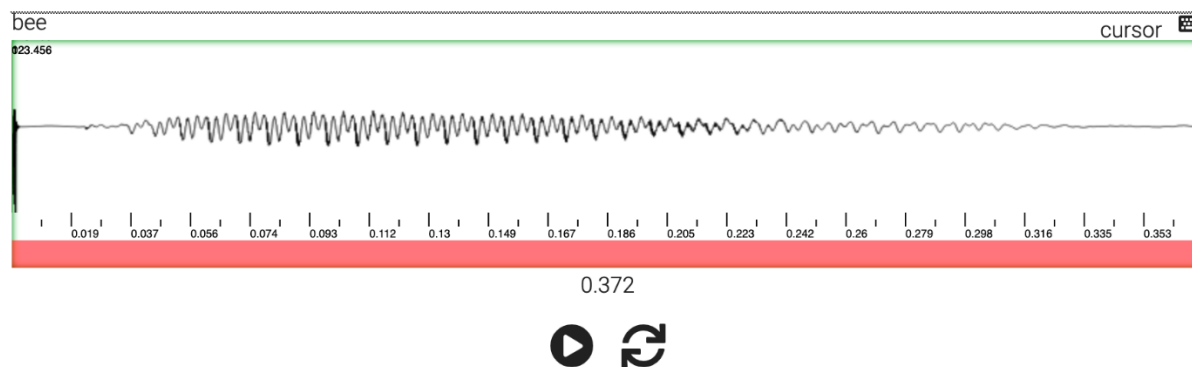
Report

Portions © 2019 University of Pennsylvania  
Funded in part by a grant from the National  
Science Foundation

ABOUT US  
FAQ  
TERMS OF SERVICE  
CONTACT

f in

For the words that your child understands please click "Understands". For the words that your child understands and also says please click "U/Says". For the words your child does not yet say or understand, please click "Neither". If your child uses a different pronunciation of a word, mark the word anyway, (e.g. 'bickie' for biscuit or 'telly' for television).



English bee

Xhosa inyosi

<b>Start</b>		<b>Stop</b>	
<b>Understands</b>	<b>U/Says</b>	<b>Neither</b>	<b>Skip</b>
		<b>Report</b>	