CV Words

Oral data from Cape Verde

Fernanda Pratas

October 2015

Contents

1. Introduction

CV Words has been created with Open Source technologies within the research project in Linguistics 'Events and Subevents in Capeverdean' (PTDC/CLE-LIN/103334/2008). This project, which has covered a three year period (2010-2013), was based in Centro de Linguística da Universidade Nova de Lisboa (CLUNL-FCSH) and had an independent funding by Fundação para a Ciência e a Tecnologia (FCT).

Currently, **CV Words** is also sponsored by FCT, through the Centro de Linguística da Universidade de Lisboa (CLUL-FLUL).

A large corpus of oral data is in constant process of semi-orthographic transcription, POS tagging and syntactic classification. The informal interviews with Capeverdean speakers are divided into discourse segments, the bigger units within CV Words. **Note: discourse segments are still called "sentences" in the current version of the database (this will soon be corrected); in the following pages of this Manual, we will consistently refer to them already as "segments".** For each segment, the following informations are also provided:

- a gloss, following the *The Leipzig Glossing Rules*
- two translations: Portuguese and English
- an audio file
- the relevant information about the speaker

Two types of searches are made available to the wide public: a basic, "Google-like" search and an Advanced search; under the latter, several elements may be combined in order to obtain precise results according to more specific needs. This is explained in detail in the following sections of this manual.

In a first phase, CV Words will include 500 **discourse segments** in a total of about 2.500 clauses and 30.000 words, resulting from 28 interviews in Cape Verde, mostly with speakers from the Santiago Island.

This corpus will soon be extended, including more data on dialectal variation across the islands of the archipelago, which is part of the scientific goals of the Dialectology and Diachrony Group, within CLUL: http://www.clul.ul.pt/

Direction, POS tagging and syntactic classification: **Fernanda Pratas**
Interviews and orthographic transcription: **Helderyse Rendall**
Web development: Solid Angle http://solidangle.eu
Events and Subevents team: **Ana Josefa Cardoso, João Costa, Luís Filipe Cunha, Alexandra Fiéis, Maria Lobo, Ana Madeira**

Email address for comments, suggestions and questions: fcpratas@gmail.com

Citation
Pratas, Fernanda. 2012. CV Words: oral data from Cape Verde. URL: http://cvwords.org/

2. Search tools

2.1. “Google-like” search

2.1.1. Purpose

This basic type of search is adequate to find naturally uttered discourse segments that illustrate the contexts in which specific **Capeverdean words**[1] may occur.

Note that CV Words is not a dictionary. Therefore, it is very probable that some words, although possibly common in this natural language, are not registered here – simply because they do not take part in any of the segments/interviews included so far. In the case you insert one of these words in the Basic search window, you will get the note: ‘No results were found’.

2.1.2. Procedure

After choosing the option Basic search, the word to be searched must be inserted in the window. This word must have at least 2 letters. Then click on the search button.

As an example, if we insert the word *ka*, which marks sentential negation in Capeverdean, the results will include all the segments that have this word, each of which will appear as follows (note again that, in the near future, ‘Sentence ID’ will be corrected to ‘Segment ID’:

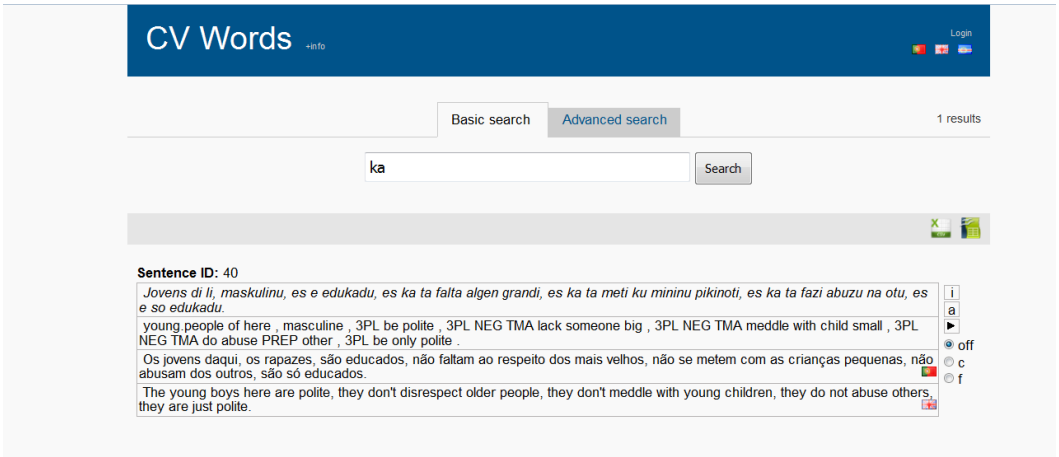

Figure A: example of result with the basic search, using the Capeverdean word *ka*.

2.1.3. Some details

(a) if more than 1 word is inserted, the results will include the segments that contain one of these words AND/OR the other, not only the segments that contain both (to search for specific combinations in each segment, one must use the Advanced search tool; see 2.2.).

[1] When introducing certain expressions in this manual, the following types have been used:
-- **boldface** expressions within the text refer to specific search objectives;
-- underlined expressions within the text refer to the buttons/shortcuts in the digital tool.
Moreover, *italic* type is always used here to write expressions in Capeverdean.

(b) since each post-verbal TMA affix is tagged morphologically as a separate word (see section 3. for the Classification decisions), if one inserts an inflected verb in the Basic search window this functions as a 2 word search, in which case the type of result obtained is the one mentioned in (a).

(c) the Basic search tool can also be used to obtain a segment for which we already know the ID number; for this, insert this simple number in the window and click on the search button (note that, during revision operations of the database, some segments may have been deleted; this means that some rare numbers may have no correspondent segment).

2.2. Advanced search

2.2.1. Purpose

This type of search is aimed at finding specific **grammatical structures** and **combinations**. Therefore, it requires a certain familiarity with the classification decisions (see section 3.) and, thus, the tags that correspond to specific desired results.

2.2.2. Procedure

The Advanced search tool combines 3 elements: **Clause**, **Syntactic function**, **Word**:

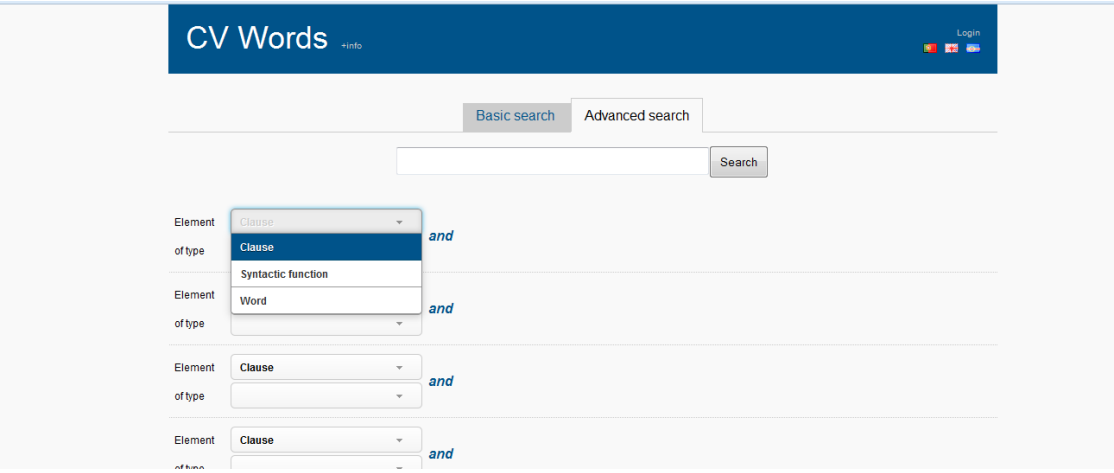

Figure B: illustration of the three elements that can be combined in the advanced search.

Each of these elements is then classified with one type tag from a closed list. These tags are abbreviations that represent expressions from the area of linguistic studies, and they have been chosen in order to make some sense in the three working languages of CV Words. There is a specific list of type tags for each element, which are as follows.

POS tags: Adj, Adv, C, Con, Det, Foc, GrIni, GrFin, H, Intj, Loc, N, Neg, Num, Pron, Pron2, PronCl, Pron, Prep, Q, Quant, Rel, TMApostV, TMApreV, V, Vaux, Vmod, Nul, Ot

Syntactic function tags: PassAg, Dir, Ind, Modif, ObjPred, Obl, Su, SuPred, Voc, Ot

Clause tags: Adj, Adv, Coord, N, Ot

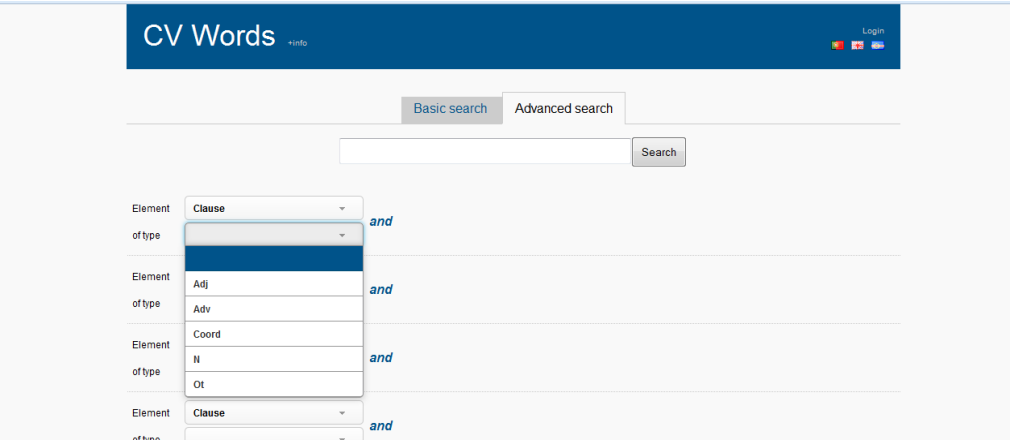

Figure C: the closed list of types for the element **Clause**.

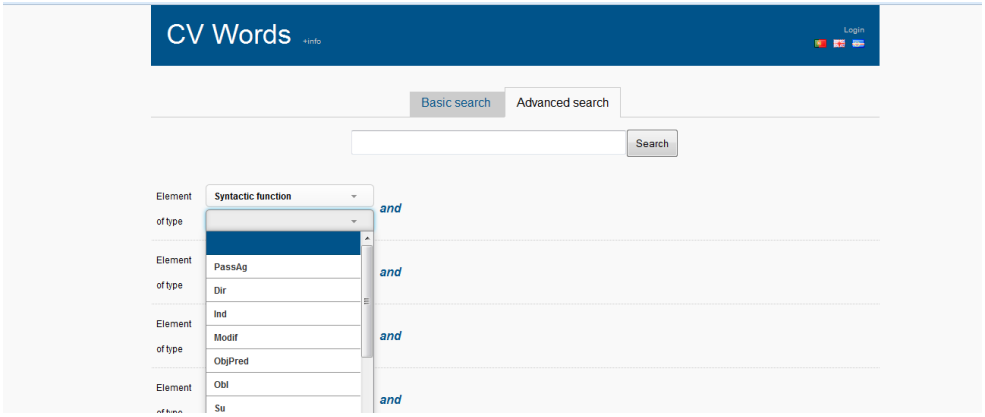

Figure D: part of the closed list of types for the element **Syntactic function**.

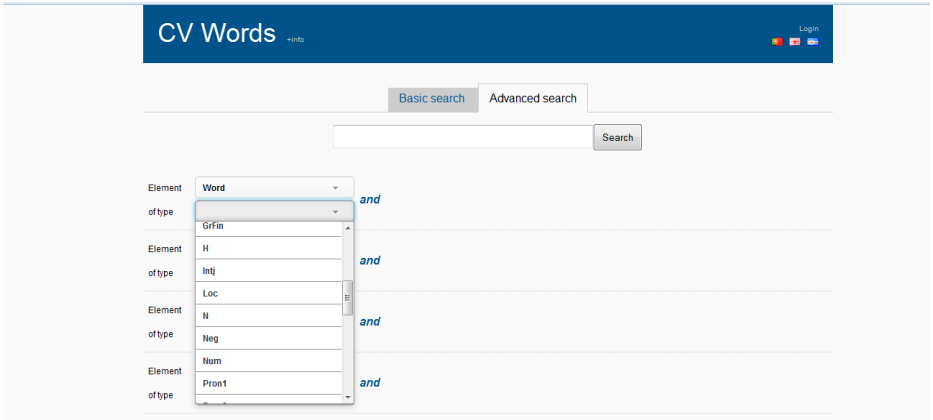

Figure E: part of the closed list of types for the element **Word**.

As said above, in section 3 these lists of tags available for each element are explained. For now, what is relevant is that one may want to find, for instance, a certain type of **Clause** that contains a certain type of **Syntactic function** (e.g. adverbial clauses that contain subjects), or a certain type of **Clause** that contains a certain type of **Word** (e.g. adjective clauses that contain relative words). Note that, obviously, this 'contain' option is

only available for the two levels above Word – that is, Clause and Syntactic function. Note, furthermore, that it is also possible to use a recursive sequence. For instance: one may want to find a certain type of **Syntactic function** that contains a certain type of **Syntactic function** (e.g. subjects that, because they are sentential subjects, contain direct objects, or indirect objects, or, say, other subjects); in the same vein, one may want to find a certain type of **Clause** that contains a certain type of **Clause** (e.g. coordinate clauses that contain relative clauses, or any other).

In the Advanced search area, one must follow this order of actions:

(a) make the choices regarding the first level you want to consider; these are done on the column that appears by default after clicking on the Advanced search button; for instance, choose Clause and then the type of Clause.

(b) if you are interested only in, say, adjective clauses (independently of what elements they may contain), just choose the element Clause and, then, the type Adj (adjective); click on the search button and you will obtain all the segments that have adjective clauses. Do not click on the + button. This is illustrated bellow:

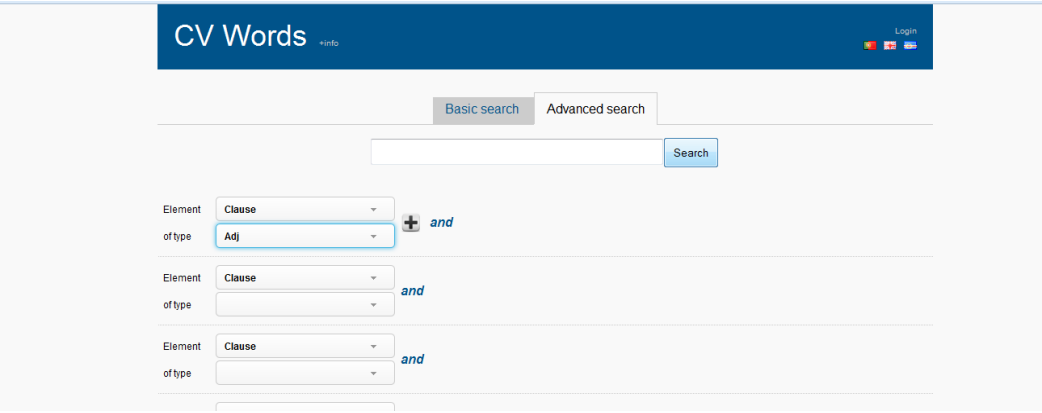

Figure F: example of a simple choice in the advanced search; this means that we want to obtain all the segments that have adjective clauses.

(c) The options at this first level that appears by default in the Advanced search area may be combined. You may want to obtain, for instance, all the segments that have preverbal TMA markers AND also postverbal TMA markers. Again, do not click on the + button. This is illustrated below:

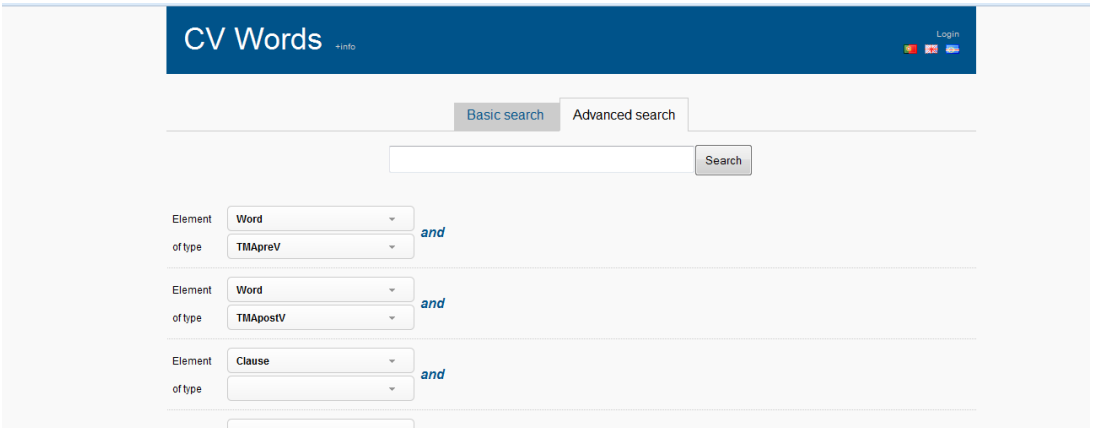

Figure G: example of a combined choice in the advanced search; this means that we want to obtain all the segments that contain preverbal TMA markers *AND* postverbal TMA markers.

(d) If, rather, you are interested in some specific Clause or Syntactic function type that, on its turn, contains another specific element type, then proceed this way: after making the first choice (on the column that appears by default), you must click on the + sign and use the extra columns to make these combinations. For instance, you want adjective clauses that contain the Syntactic function subject and also a relative Word: on the column that appears by default, choose the element Clause of the type Adj (adjective); then click on the + sign; by doing this, you make two other columns appear; then, in the next column, choose the element Syntactic function of the type Su (subject), and in the last column choose the element Word of the type Rel. Then click on the search button. You will obtain all the adjective clauses that contain subjects AND relative words.

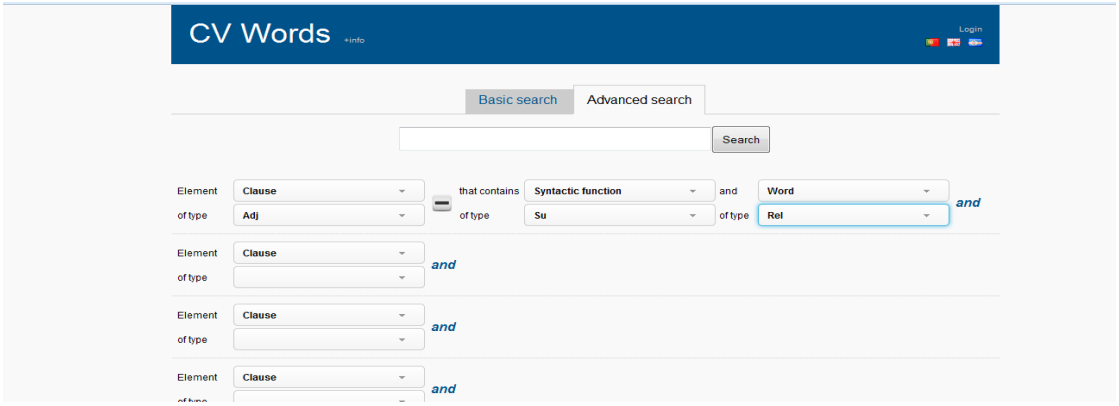

Figure H: illustration of the possible combinations in the advanced search.

2.2.3. Some details

Note that, in the option described in (d), the elements combined by AND may either be in different parts of the segment or one of them may be within the other. In other words, in the search exemplified in Figure H, the relative words are not necessarily within the subjects.

3. Classification decisions

The classifications available at this point of the database are presented in this section. The three first subsections show the respective topic closed lists, with some brief notes explaining those of them that seem to need some kind of clarification. For others, whose motivation seems more obvious, no specific account is provided. The fourth section, about the glosses, also explains some decisions.

All these decisions have been taken after several months of thorough discussion and reflections, which considered also other annotation systems for other languages and several tests using working hypotheses. More explanations and details that may prove to be necessary in the near future will be added in further versions of this manual.

3.1. Parts-of-speech (POS) tagging

The complete list of POS tags and their respective abbreviations is below:

Tag	meaning
Adj	adjective
Adv	adverb
C	(is applied to the word *ki* in certain contexts, like clefts)
Con	connective (is applied to all connectives, including conjunctions)
Det	determinant (applies to all determinants, including demonstratives)
Foc	focus
GrIni	initial graphic sign (is applied to the opening parentheses)
GrFin	final graphic sign (is applied to all closing graphic signs)
H	hyphen
Intj	interjection
Loc	locative (applies to some adverbs, part of demonstrative expressions related to degrees of proximity)
N	noun (no distinction is made between proper and common nouns)
Neg	sentential negation (specific for *ka* in Santiago; all other negative words are labelled with their category: pronoun, adverb, etc.)
Num	numeral
Pron1	strong pronoun (personal pronouns like *ami, abo*…, generally used in clitic doubling)
Pron2	free pronoun (personal pronouns: *mi, bo*…, which are 'free' as opposed to clitics)
PronCl	clitic pronoun
Pron	other pronominal form (this applies to non-personal pronouns)
Prep	preposition
Q	interrogative word
Quant	quantifier
Rel	relative word
TMApostV	post-verbal Tense, Mood, Aspect morpheme (applies to morphemes that are affixed to the verb; although they appear as a part of the verb, they are tagged here separately, so that a search can be focused on them; this is motivated by the fact that the original main drive of this database has been the study of tense, mood and aspect in Capeverdean)
TMApreV	pre-verbal TMA (applies to TMA morphemes that appear in preverbal position)
V	main verb
Vaux	auxiliary verb
Vmod	modal verb
Nul	null category
Ot	other (planned to fulfil any needs not yet predicted in this first version of the database)

3.2. Syntactic functions

The complete list of syntactic functions and their respective abbreviations is below:

Tag	meaning
PassAg	agent complement in passives
Dir	direct object
Ind	indirect object
Modif	modifier
ObjPred	object complement
Obl	oblique object
Su	subject
SuPred	subject complement
Voc	vocative
Ot	other (planned to fulfil any needs not yet predicted in this first version of the database)

3.3. Clauses

The complete list of clauses and their respective abbreviations is below:

Tag	meaning
Adj	adjective clause
Adv	adverbial clause
Coord	coordinate clause
N	noun clause
Ot	other (in some cases, it is used with what would be considered the main clause)

3.4. The glosses

As said in the Introduction, the English glosses in CV Words follow the *The Leipzig Glossing Rules*: http://www.eva.mpg.de/lingua/resources/glossing-rules.php

Some small adaptations have been made, in order to fulfil specific language requirements, but, hopefully, these are very clear.

3.4.1. Relation between the glosses and the POS tags

In many cases, the gloss is identical to the POS tag abbreviation (except for the letter case – upper case for glosses, capitalisation for POS abbreviations: e.g. PREP / Prep, for preposition). This is intentional, since the POS tags are only available online through the tooltips (see section 5), not when we export the results to a CSV or to a ODF file. Once in these files, the only information about the meanings and grammatical properties of individual words and parts of words is available through the English glosses.

3.4.2. Distinction between the period and the colon

Period

A. When the meaning of one Capeverdean word or morpheme is translated into English in more than one word or morpheme, a period is used to separate these different units:

Capeverdean word	→	**English gloss**
bisnetu		great.grandchildren

B. Also, when one morpheme carries two different grammatical informations, a period is used within the gloss to separate the representation of these informations. For instance, the morpheme *-da*, which marks passive and past:

Capeverdean morpheme	→	**English gloss**
-da		PASS.PST

Colon

When there are two different morphemes in one word (thus, two different meanings must be glossed), a colon is used to separate these units. For instance, the verb *kume* 'eat' and the morpheme *-ba*, which marks past:

Capeverdean word	→	English gloss
kumeba		eat:PST

3.4.3. Irregular past verbs

Some Capeverdean verbs may dispense with the use of *-ba* to mark past, using instead a form that is borrowed from Portuguese, the European lexifier of this Creole language. For instance *sabia* instead of *sabeba*, meaning 'used to know'. In these cases, the gloss does not provide separate meanings for the verb root and the past morpheme (since they are not separable), but it directly gives the past meaning. This is as follows:

Capeverdean word	→	**English gloss**
sabeba		know:PST
sabia		used.to.know

3.5. Transcription rules

The segments produced by the Capeverdean speakers in the oral interviews are subject to semi-orthographic transcription, using the tool Elan (http://tla.mpi.nl/tools/tla-tools/elan/elan-description/) and following ALUPEC, the Capeverdean official alphabet.

The orthography of every word has been normalized, independently of the way in which it is pronounced by the speakers. For this, the *Dicionário do Crioulo de Santiago (Cabo Verde)* (Brüser and Santos 2002, organized under the direction of Jürgen Lang) has been used as reference. There have been, however, some small adaptations, like the elimination of all accents.

Other important notes about the transcription are:

-- The hesitations of the speakers are marked by three periods and the parenthetical segments of speech are indicated by parentheses.

-- Since brackets are not available here, clarification notes added by the transcriber are indicated by double parentheses, like in the following segment:

Tudu, N ta nkaminhaba es pa PMI ((Programa Materno Infantil)).

-- Discourse segments that are not clear and, thus, have not been transcribed, are indicated by three periods inside double parentheses, like in the following segment:

Ben more-m un rapas la San Martinhu ki e nha subrinhu, fidju di nha subrinha((...)).

4. Information about the Capeverdean speakers

4.1. Education degrees

Tag	meaning
Nul	no official level of education
Bas	basic school
Sec	high-school
Sup	university

4.2. Counties in Cape Verde

Boa Vista, Brava, Maio, Mosteiros, Paul, Porto Novo, Praia, Ribeira Brava, Ribeira Grande de Santiago, Ribeira Grande de Santo Antão, Sal, São Domingos, São Filipe, São Lourenço dos Órgãos, São Miguel, São Salvador do Mundo, São Vicente, Santa Catarina do Fogo, Santa, Catarina de Santiago, Santa Cruz, Tarrafal de Santiago, Tarrafal de São Nicolau.

5. Visualization of the results

The results will appear in a list of segments, all visible as is illustrated in Figure A (p. 4). All the information available for each segment can be viewed in a very simple and friendly way.

5.1. POS tagging

When we put the cursor over a Capeverdean word, a tooltip appears that indicates its POS tagging – e.g. *inda* 'yet' is an adverb. See this below:

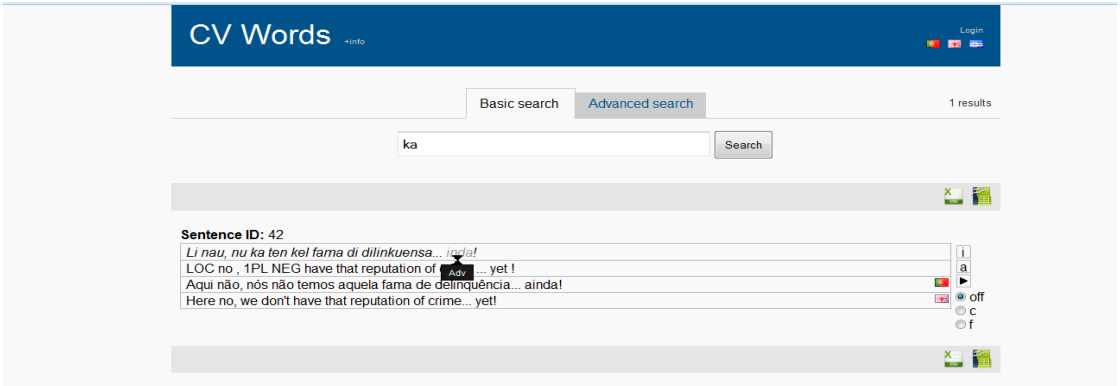

Figure I: visualization of a POS tag, when we put the cursor over a word on the 'results' page.

5.2. Syntactic functions

When we select the 'f' button, on the right, the different syntactic functions in the segment are shown inside the following signs (*….*) (this is not to be confused with the true parentheses, which mark parenthetical segments in the speakers' discourse). With the cursor over this sign, a tooltip appears that indicates the type of the syntactic function. Each syntactic function belongs to a specific clause within the segment; this clause id is also indicated.

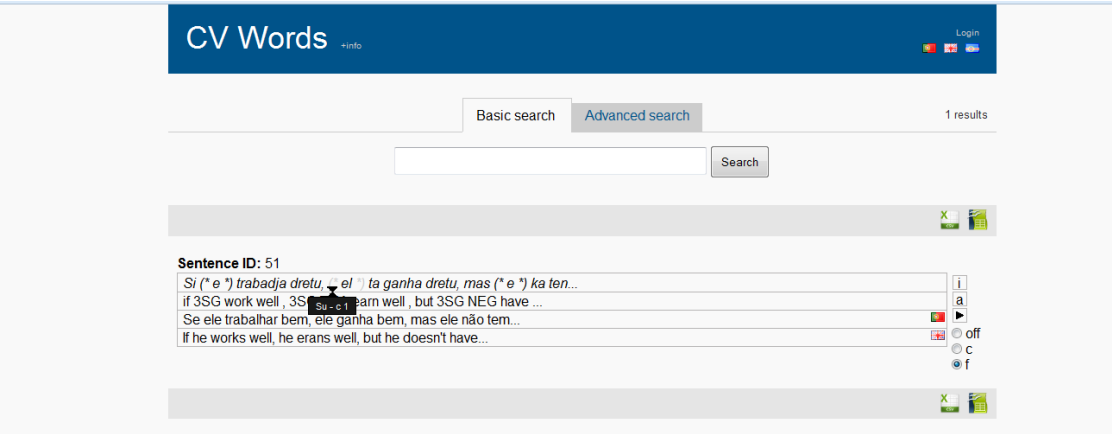

Figure J: visualization of a subject that belongs to the clause whose id is 1.

5.3. Clauses

When we select the 'c' button, on the right, the different clauses within the segment are shown inside the following signs (*….*) (this, again, is not to be confused with the true parentheses, which mark parenthetical segments in the speakers' discourse). With the cursor over this sign, a tooltip appears that indicates the clause type. Each clause has its own id within the segment.

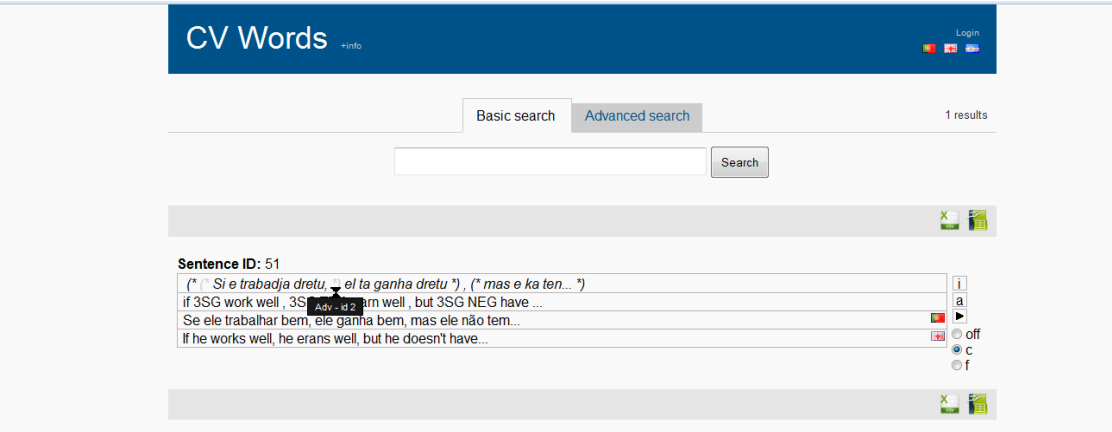

Figure K: visualization of an adverbial clause, whose id within the segment is 2.

5.4. Audio file and information on the speaker

To listen to the audio file of each segment, click on the ► button.

To know the details about the speaker (name, origin, level of education and age, click on the **a** button. This is illustrated below.

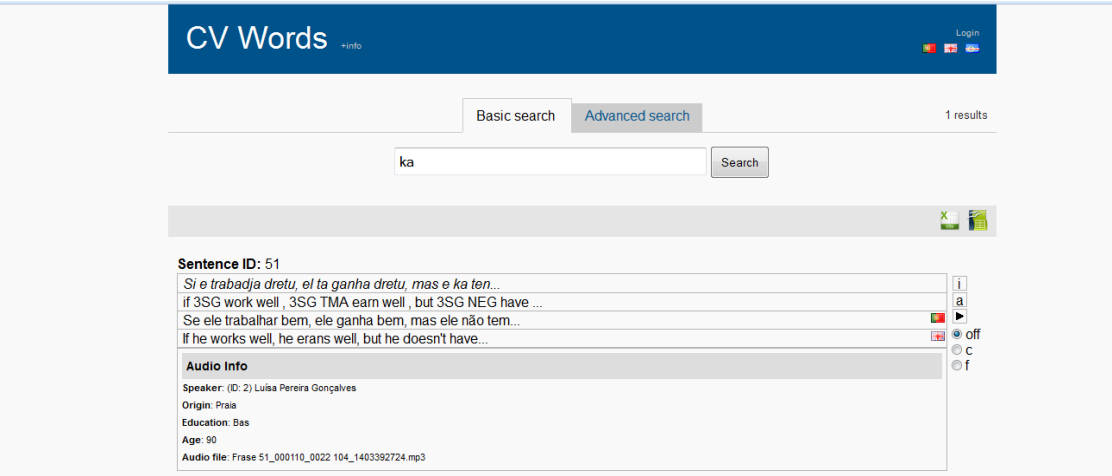

Figure L: visualization of the information about the speaker.
