Phonemic Tagging with the G2P BAS Web Service

The Bavarian Archive for Speech Signals (BAS), has kindly published a set of speech processing web services including one for phonemic transcription called G2P. You can use this service yourself directly, using your web browser, but LaBB-CAT also has a module for using it automatically, called the *BAS Services Manager*.

> **Warning**
>
> In order to function, your LaBB-CAT server must be able to connect to the internet.

> **Caution**
>
> Using G2P for phonological tagging requires LaBB-CAT to send your orthographic transcripts over the internet to a third party. Although point 3 of the BAS Web Services Terms of Service makes clear that uploaded data is deleted after 24 hours, using the service is only suitable in situations in which you have consent from participants to do so.

You can use G2P for phonological tagging if your speech is in any of the following languages:

- Albanian
- Australian Aboriginal Languages
- Afrikaans
- Albanian
- Basque
- Catalan
- Dutch
- English
- Estonian
- Finnish
- French
- Georgian
- German
- Hungarian
- Italian
- Japanese
- Kunwinjku
- Luxembourgish
- Maltese
- Norwegian
- Polish
- Romanian
- Russian

- Spanish
- Swedish
- Yolŋu Matha

LaBB-CAT must be able to identify which language each transcript is in, so you must ensure the language is set either

- in the transcript's Language transcript attribute, or
- on the corpora page (where you can define the language for all transcripts each corpus).

The available language options can be set in LaBB-CAT by going to the transcript attributes page and clicking the Options button of the "language" attribute. The value must be a two-letter ISO639-1 code optionally appended with a two-letter country code - e.g. "en" or "en-NZ".

Install the layer manager

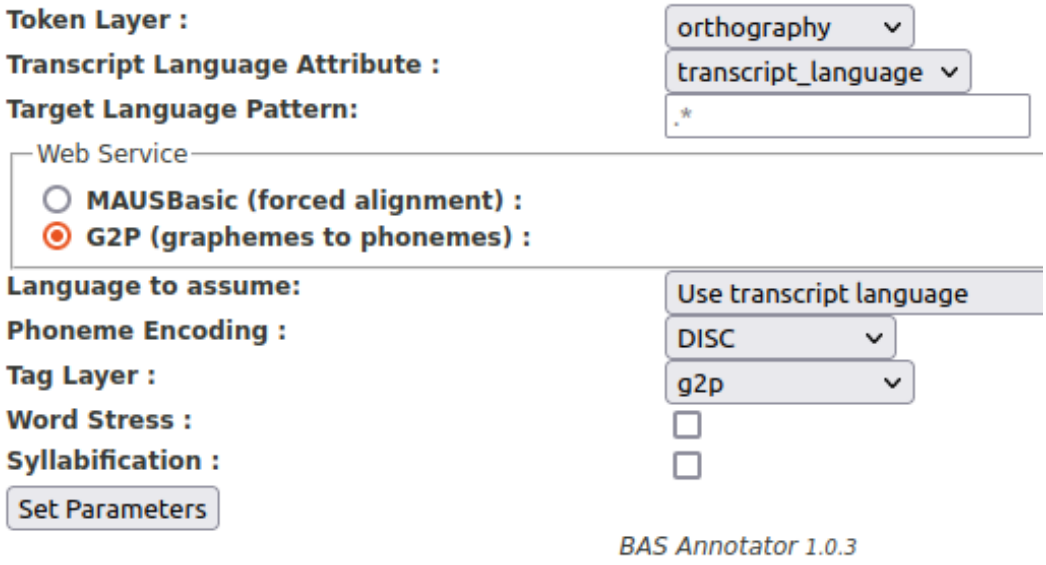
1. In LaBB-CAT, select the *layer managers* option on the menu, which gives you a list of the layer managers already installed.
2. At the bottom of the page, follow the *List of layer managers that are not yet installed* link.
3. Look for *BAS Web Services Manager* in the list, and press it's *Install* button, and then *Install* again.
4. Follow the "terms of usage" link and read the terms.
5. Close the terms page, returning to LaBB-CAT.
6. Select true for the "Accept Terms of Usage" option
7. Press *Configure*.
   You will see a page of information about the Layer Manager, including instructions on how to set up forced alignment.

Using G2P for phonemic transcription

1. Select the *word layers* option on the menu - this will display a list of all the word layers you already have in the database.
2. At the top of the list, there's a blank form for creating a new layer - fill this form in:
   - **Layer ID**: enter something like `phonemes`
   - **Type**: select *Phonological* (Or *Text* if you don't want to use "DISC" encoding; see below)
   - **Alignment**: select *None* (as these are simply tags on the orthographic words)
   - **Manager**: select *BAS Web Services Manager*
   - **Generate**: select *Always*
   - **Description**: *Phonemic transcription according to BAS Web Service's G2P*

3. Press the *New* button to create the layer.
   You will see a form that allows you to configure the layer; check the online help for that page to guide you.
   Options include:

   - **G2P** - ensure you select this option for phonemic transcription.
   - **Phoneme Encoding** - the encoding of the phonemes, which includes all of the options supported by G2P, plus "DISC" which, if selected, invokes G2P with "sampa" as the encoding option, and then converts the result to CELEX's DISC encoding, which uses exactly one character per phoneme. The "DISC" option is recommended if the layer has its type set to "phonological".
   - **Word Stress** - prefix stressed vowels with a stress marker
   - **Syllabification** - include syllable boundary markers in the transcriptions.



4. Press *Set Parameters*
5. Press *Regenerate*

LaBB-CAT will then generate annotations for all the transcripts you already have in your database. If you have a lot of data, this may take a while.

From now on, when you upload a new transcript, the G2P annotations will automatically be generated for it.