
# Hinweise für die Anfertigung von Transkriptionen

Das ASD-Corpus umfasst 2274 Audiodateien, deren Inhalt nur durch Transkription erschlossen werden kann. Für die Erfassung historischer und/oder ethnographischer Information ist eine orthographische und standardnahe Transkription im allgemeinen ausreichend. Die sprachwissenschaftliche Analyse des Materials hingegen kommt ohne eine phonetische Transkription nicht aus. Da diese jedoch im Vergleich zu einer orthographischen Transkription wesentlich zeitraubender ist und überdies nur von Spezialisten durchgeführt werden kann, soll die Erschließung des ASD-Corpus durch die simultane Anfertigung von beiderlei Transkriptionsarten erfolgen. Die orthographische Transkription kann von sprachwissenschaftlichen Laien angefertigt werden. Voraussetzung ist jedoch, dass diese den siebenbürgisch-sächsischen Dialekt verstehen.

## *Regeln*

Im Normalfall werden die Transkriptionen - sowohl die orthographischen wie auch die phonetischen - unter Verwendung des kostenlos verfügbaren Programmes "**Praat**" transkribiert (http://www.fon.hum.uva.nl/praat/). Eine ausführliche Anleitung zur Verwendung von Praat findet sich weiter unten in Anhang II.

Für die Transkription dürfen *ausschließlich* die folgenden Zeichen verwendet werden:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Alle anderen Zeichen können bei der elektronischen Datenverarbeitung zu mehr oder weniger großen Problemen führen. Das gilt auch für die **deutschen Umlaute** und das scharfe "s" (ß). Diese müssen durch "a#", "o#", "u#" und "s#" wiedergegeben werden.

Für die **phonetische Transkription** sind von den Autoren des Programmes Praat Regeln festgelegt worden, wie die Zeichen des phonetischen Alphabets unter ausschließlicher Verwendung der oben gelisteten Zeichen wiederzugeben sind (s. Anhang II).

Bei der **orthographischen und standardnahen Transkription** ist Folgendes zu beachten: die Morphologie wird dem Standard angepasst, **dialektale Formen** werden jedoch in spitzen Klammern erhalten.
Wenn sie sich auf Einzelwörter beziehen, werden sie *nachgestellt*:
*gewesen<gewest>*
*der<das> Saal*
*wahrscheinlich<ken sei>*.

Bei Mehrwortausdrücken und phraseologischen Wendungen wird der dialektale Ausdruck in spitzen Klammern als öffnendes Glied *vorangestellt*, durch den Standardausdruck wiedergegeben und am Ende des Standardausdrucks mit </> abgeschlossen:
*<in die Rent kommen>ins Lot kommen</>*
*<e broien>zu brennen</>*
*<ister>auf einmal</>*

**Fremdsprachige Passagen** werden wie folgt markiert:

*<f>fremdsprachige Passage</f>*

Dabei werden für *f* folgende Sprachkürzel eingesetzt:

r = rumänisch; u=ungarisch; rus=russisch; ukr=ukrainisch; d=hochdeutsch;
jid=jiddisch; engl=englisch;

Die deutsche Übersetzung längerer fremdsprachiger Passagen wird in die Anmerkungsdatei geschrieben.

Phonetische Sonderzeichen sind zwischen spitzen Klammern zur Not möglich, sollten jedoch vermieden werden. Sie müssen als ASCII-Zeichen gemäß den Praat-Regeln (s.o.) eingegeben werden.
Sofern möglich, müssen die Wörter gemäß der neuen deutschen Rechtschreibung eingegeben werden.
Fremdsprachige Wörter sollen, sofern sie erkannt werden, in der für diese Sprache üblichen Schreibweise eingegeben werden. Nach Möglichkeit soll die Vermittlersprache, nicht die Ursprungssprache (Sprache des Etymons) angegeben werden. Wenn z.B. ein ungarisches Wort über das Rumänische ins Siebenbürgisch-Sächsische gekommen ist, wird das Wort als rumänisch gekennzeichnet.
Wörter, die nicht identifiziert werden können, sollen in naheliegender Orthographie - bei strikter Vermeidung von Sonderzeichen! - den Höreindruck wiedergeben.

Beispiel:
Das Wort "zseb" (ungar. für "Hosentasche")
- wird mit "Zseb" transkribiert, wenn es vom Transkriptor als ungarisch erkannt werden würde
- könnte z.B. mit "Scheb" transkribiert werden, wenn es nicht erkannt werden würde.

Einzelwörter, die als fremdsprachlich identifiziert werden, werden folgendermaßen notiert:
*Buchhalter<r>contabil</r>*
*Kulturheim<r>camin cultural</r>*

Bei Mehrwortausdrücken wird der fremdsprachliche Ausdruck vorangestellt:
*<<r>au venit dupa noi</r>>haben uns abgeholt</>*

**Probleme aller Art** werden auf folgende Weise gekennzeichnet: {?Notiz(en)}
Wenn Äußerungen in die Notizen eingebaut werden, sind diese vom Kommentar durch Doppelpunkte abzusetzen. Dialektausdrücke werden wie üblich in spitze Klammern gesetzt.
*{?unversta#ndlich}*
*{?Stimme aus dem Hintergrund :auch drei:}*

*{?:<Zeach>:}*

Aus Gründen des Datenschutzes dürfen Eigennamen von Personen ***nicht transkribiert*** werden. An ihrer Stelle sind drei Gitterkreuze (###) zu schreiben.

**Satzzeichen** werden in den Äußerungen grundsätzlich *__nicht__* verwendet!
(Ausnahmen in geklammerten Notizen sind möglich)

**Anmerkungsdateien**

Erklärungen, z.B. Bedeutungsangaben (*Christtag* → Anm. ‚Weihnachten') müssen in einer gesonderten Datei abgelegt werden, für die folgende Konventionen gelten:

- Jeder Anmerkung muss die zugeordnete **Intervallnummer** in eckigen Klammern und gefolgt von einem Doppelpunkt (Beispiel: [15]: ) vorangestellt werden. Mehrzeilen-Bereiche werden so notiert: [15-22]:
- In der ersten Zeile der Datei muss der Dateiname stehen (Beispiel: „Datei 1445b-05").
- In der zweiten Zeile der Datei wird der Herkunftsort des Informanten angegeben (Beispiel: „Anmerkungen zu Keisd"). Diese **Überschrift** muss durch **eine Leerzeile** von der ersten Anmerkung getrennt sein.
- **Innerhalb einer Anmerkungseinheit dürfen keine Zeilenumbrüche vorkommen.**
- Schlagwörter zum Dateiinhalt als Ganzem werden vor den ersten intervallweisen Anmerkungen notiert (Beispiel: „Schlagwoerter: Hanklich; Hochzeit; ...").
- Am Ende der Datei, abgetrennt durch **zwei Leerzeilen**, muss der **Name des Autors** in eckigen Klammern vermerkt werden.
- Bei der Datei muss es sich um eine **reine Text-Datei in utf8-Kodierung** handeln (auf keinen Fall eine Word-Datei!).
- Außer der einen Leerzeile zwischen Überschrift und Anmerkungen und den zwei **Leerzeilen** zwischen Anmerkungen und Verfassernamen darf die Datei keine Leerzeilen enthalten.
- Der **Dateiname** muss aus der Nummer des Interviews und der Erweiterung „phon" bzw. „orth" für die Art der Transkription und der Endung „txt" bestehen (Beispiel: 1445b-05.orth.txt).

## *Herunterladen der Audio-Dateien*

Die Audiodateien des ASD können über den Menüpunkt "CORPUS" einzeln von der Webseite des Projekts heruntergeladen werden:


In der Rubrik "Audio" erscheint die Nummer der Audio-Datei. Mit dieser Nummer muss die Transkriptionsdatei benannt werden!

Das Herunterladen der Audiodatei ist erst nach Eingabe eines Passwortes möglich. Um dieses Passwort zu erhalten, schicken Sie bitte eine Mail an thomas.krefeld@lmu.de oder luecke@lmu.de. An diese Mailadressen senden Sie schließlich bitte auch die fertige Transkriptionsdatei.

## *Übersenden der Transkripte*

Die fertigen Transkripte müssen in Form von "Textgrid-Dateien" (erzeugt von "Praat", s.u.) über folgende Internetadresse hochgeladen werden:

http://www.asd.gwi.uni-muenchen.de/index.php?intern=true&upload=true

Für die Nutzung dieser Funktion benötigen Sie eine Benutzerkennung, die wir Ihnen gerne einrichten. Bitte wenden Sie sich an:

luecke@lmu.de oder thomas.krefeld@lmu.de

Grundsätzlich gilt: Die Textgrid-Dateien ***müssen*** denselben Namen wie die zugehörige Audio-Datei besitzen (Beispiel: 566b-03.mp3 => 566b-03.TextGrid)

**Bitte fertigen Sie zunächst eine Probe-Transkription an und senden Sie diese an eine der oben genannten Mailadressen!**

- **Anhang I: Das Transkriptionsprogramm "Praat"**

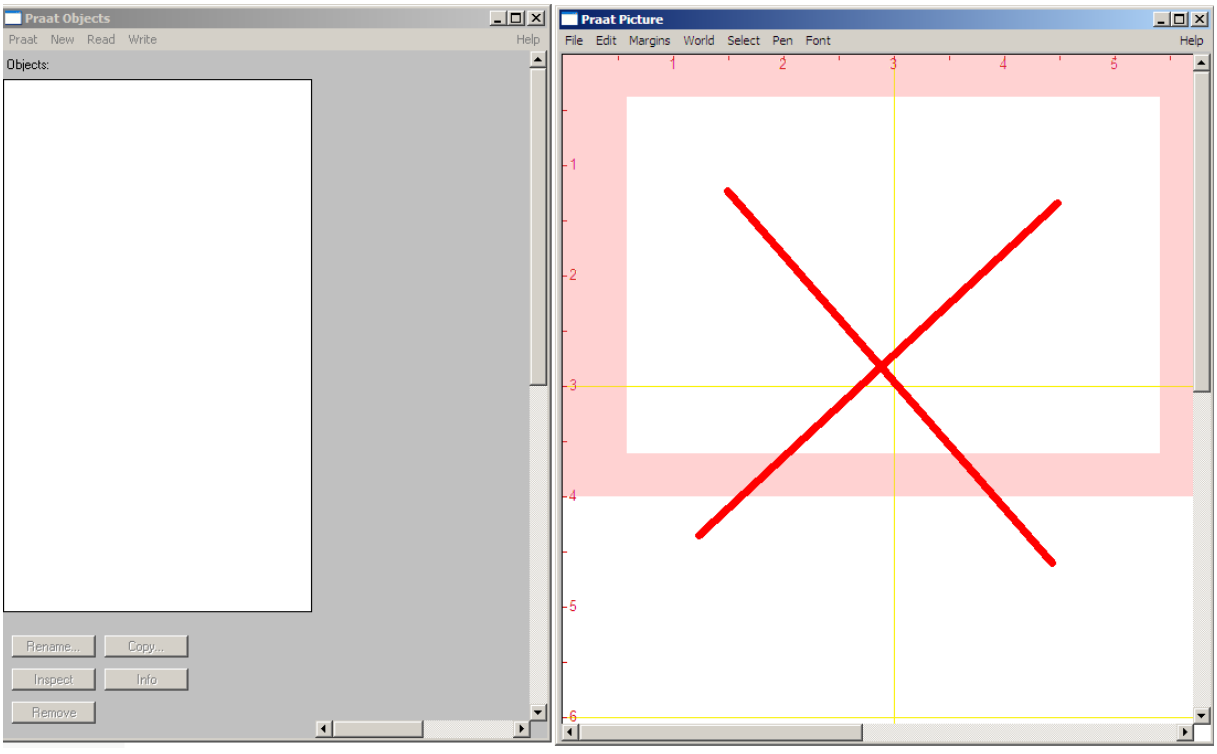
Das Programm Praat ist von Phonetikern für Phonetiker gemacht und besitzt einen beeindruckenden Funktionsumfang. Aus Sicht der Korpuslinguistik ist es vor allem deswegen interessant, weil damit Audio-Aufnahmen von Sprache transkribiert und für die weitere Verarbeitung mit Skripts und in Datenbanken aufbereitet werden können.

Praat ist kostenlos und für alle gängigen Betriebssysteme verfügbar. Es kann unter http://www.fon.hum.uva.nl/praat/ heruntergeladen werden. Praat wird ständig aktualisiert. Man sollte sich stets die neueste Version von der Webseite der Entwickler herunterladen.

Das folgende Beispiel präsentiert die Transkription einer Audioaufnahme der Wenkersätze.

0. Starten des Programms

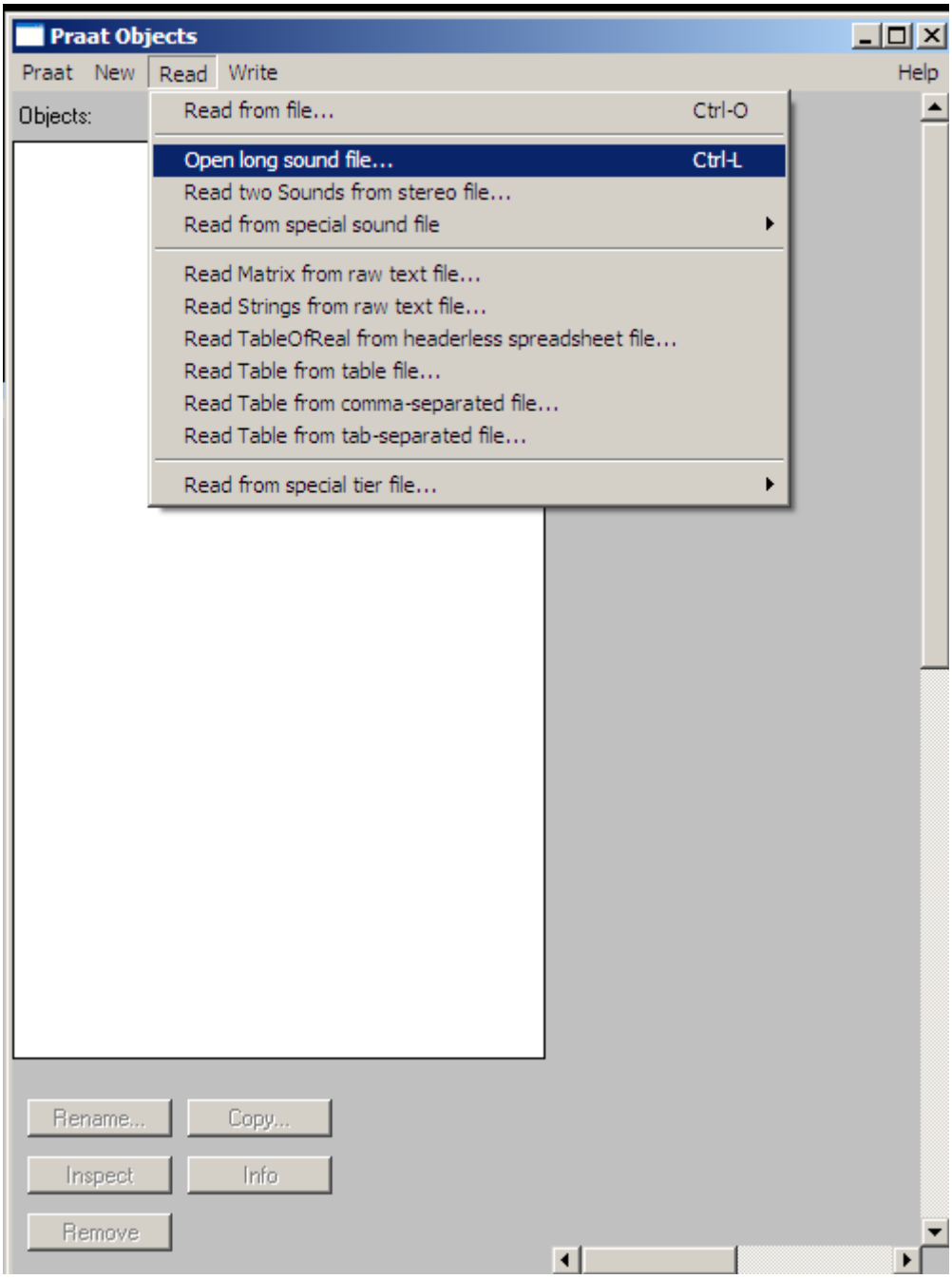
Das Programm wird durch Doppelklick gestartet. Praat öffnet automatisch zwei Fenster, nämlich "Praat Objects" und "Praat Pictures". Letzteres wird nicht benötigt und kann gleich wieder geschlossen werden.



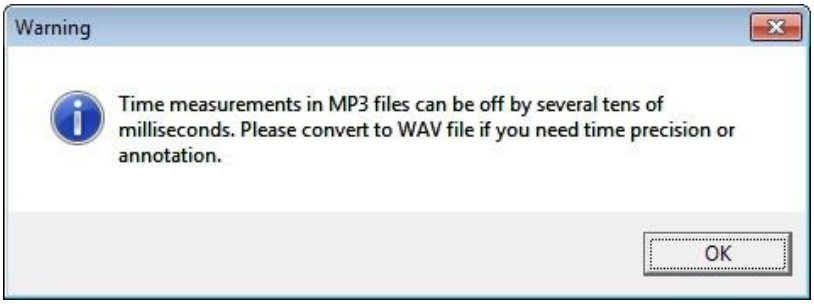

**Codierung:**
Bevor Sie mit dem Transkribieren beginnen, müssen Sie einmalig im Fenster "Praat Objects" die Einstellungen anpassen. Dazu gehen Sie oben links auf "Praat", dann auf "Preferences / Text writing preferences…" und wählen **UTF-8** aus.

1. Öffnen der Audio-Datei

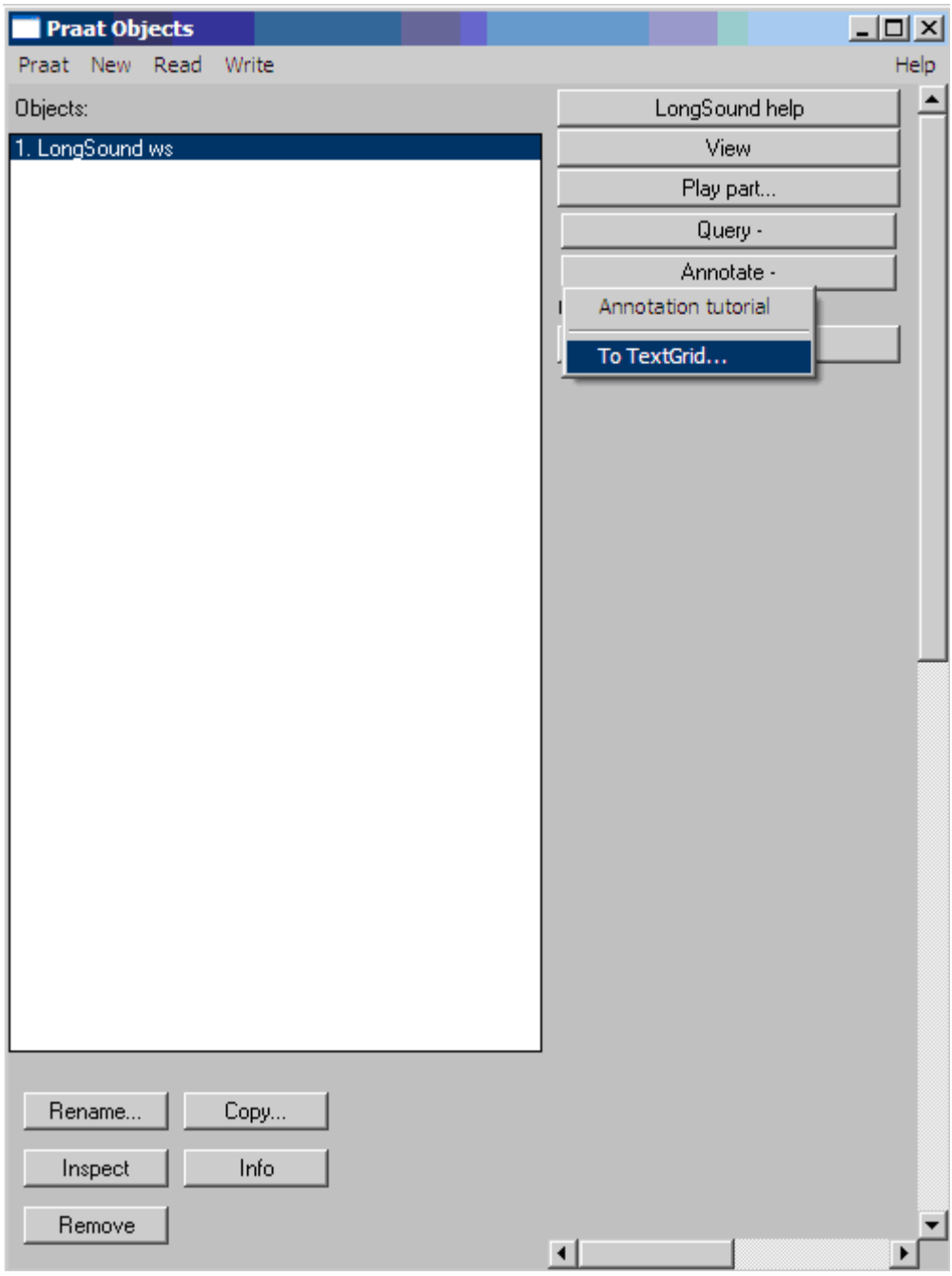
Gehen Sie auf "Read" und dann auf "Open long sound file ..."




Bei mp3-Dateien erscheint folgende Meldung, einfach OK klicken und weiter:

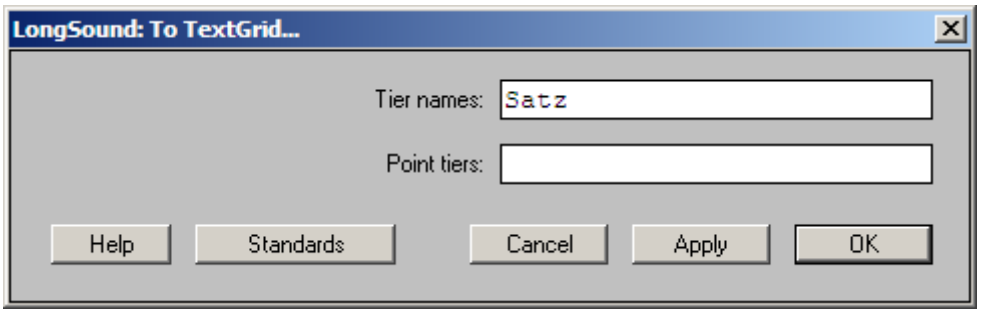



2. Markieren Sie die geöffnete Audio-Datei in der Liste der "Objects" und klicken Sie auf "Annotate" und anschließend auf "To TextGrid..."

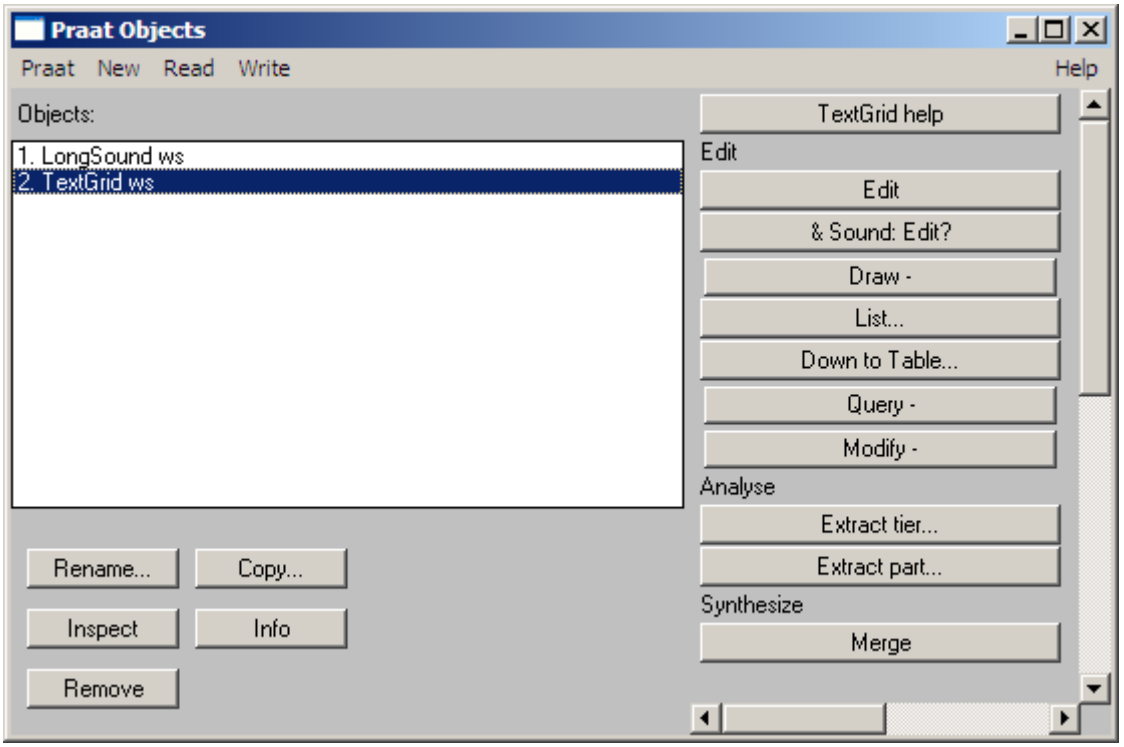


3. Praat fragt dann nach den Namen der sog. "Tiers". Damit sind zunächst Ebenen gemeint, die für die separate Transkription der Äußerungen verschiedener Sprecher gedacht sind. Im Sinne der Korpuslinguistik hat es sich in vielen Fällen als günstig erwiesen, unterschiedliche grammatische Kategorien diesen Ebenen zuzuordnen. Entsprechend würde man "Satz", "Wort" und "Silbe" als Tier-Namen wählen. Die Entscheidung ist natürlich vom jeweils verfolgten Forschungsinteresse abhängig.

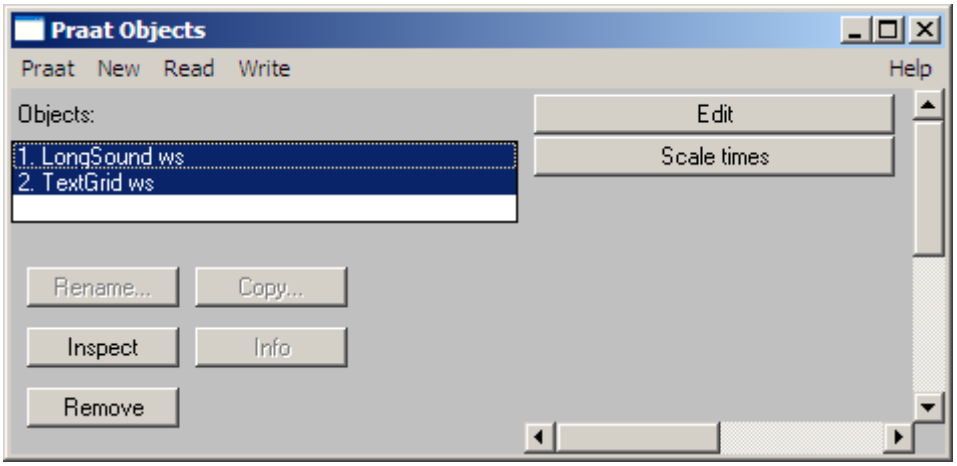
Löschen Sie alle automatisch angelegten Einträge und legen Sie ein Tier namens "Satz" an, in das der Text satzweise transkribiert wird:



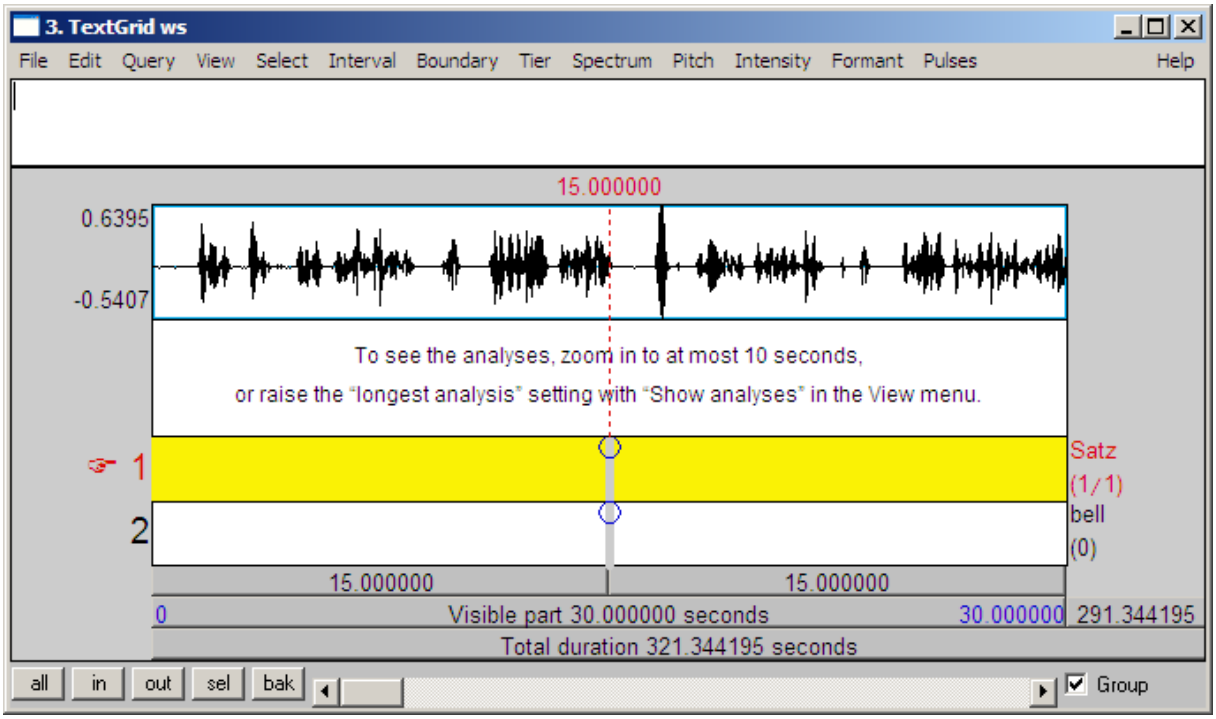

Nach dem Klick auf "OK" wird der Objektliste ein neuer Eintrag hinzugefügt, nämlich "TextGrid xx":




Nun müssen beide Listeneinträge gemeinsam markiert werden (dazu hält man beim Mausklick die Strg-/Ctrl-Taste gedrückt):



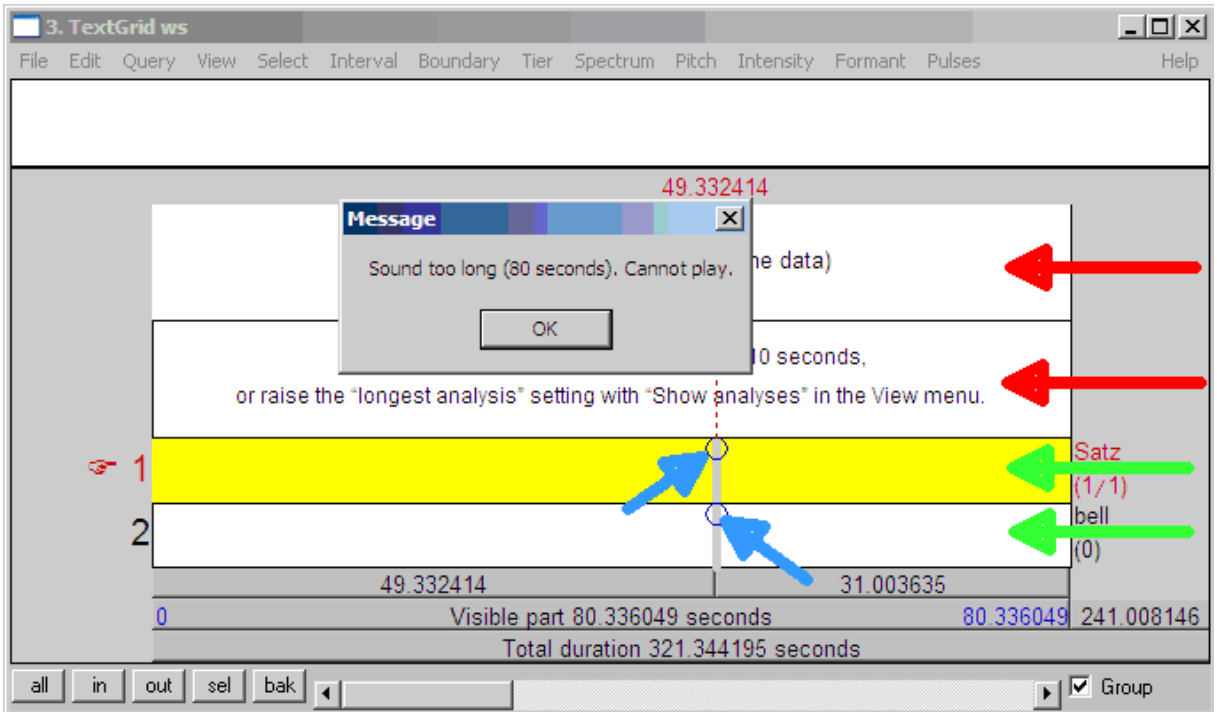

Ein Klick auf "Edit" öffnet folgende Ansicht:




Die Größe des Ausschnitts lässt sich mit den kleinen Knöpfen am linken unteren Fensterrand verändern.

Nun folgt ein wichtiger Schritt: die Segmentierung der Audiodatei in einzelne Intervalle. Auf alle Fälle sollten Passagen unterschiedlicher Sprecher in jeweils eigene Intervalle gesetzt werden. Am Anfang eines jeden Intervalls muss der Sprecher angegeben werden. Dabei wird der Informant stets mit #1# bezeichnet. Weitere Sprecher werden in der Reihenfolge ihres Auftretens mit #2#, #3# etc. bezeichnet. Sprechen mehrere Sprecher gleichzeitig, werden sie zusammengefasst: #1,2,4# oder gegebenenfalls #alle#. Ist unklar, ob es sich bei einem Sprecher um einen bereits aufgeführten oder einen neuen handelt, kann er mit #?# bezeichnet werden.

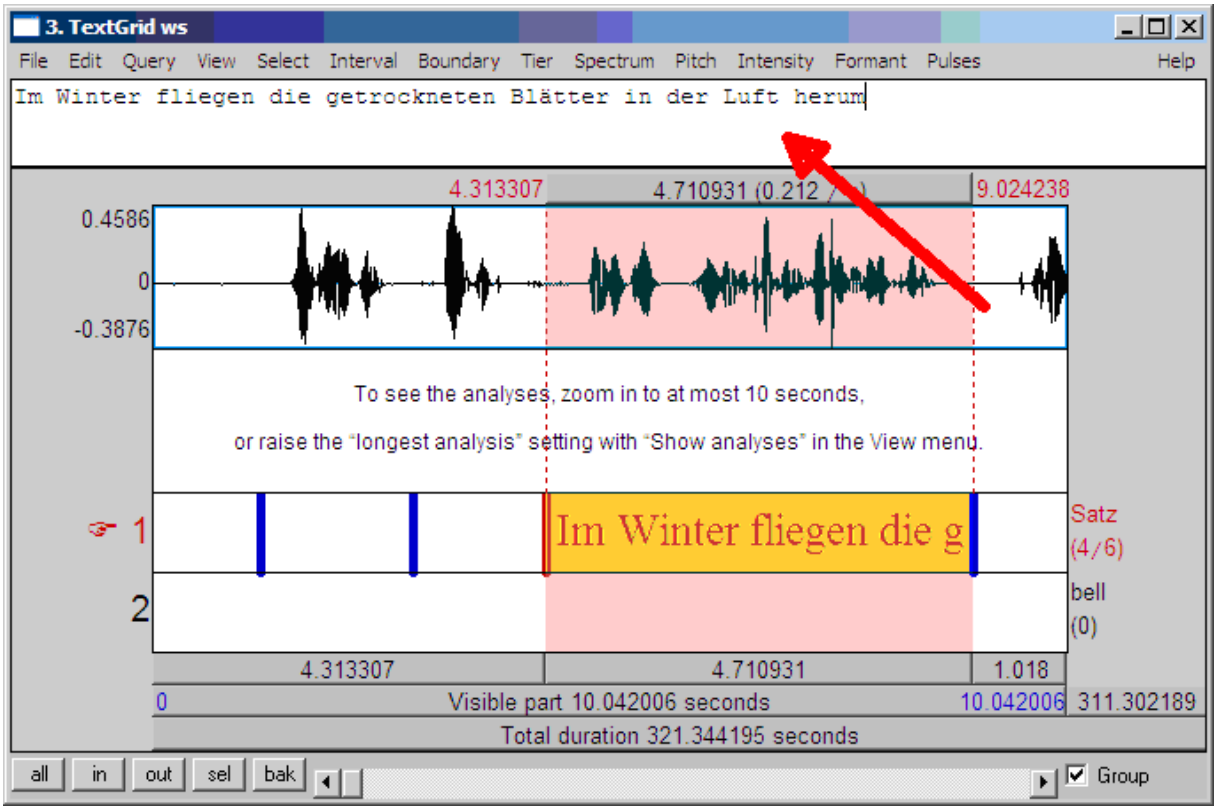
Das Abspielen des Tons wird durch Drücken der Tabulator-Taste ausgelöst. Praat spielt nur Ausschnitte von maximal 60 Sekunden Länge ab. Bei längeren Ausschnitten erfolgt eine entsprechende Fehlermeldung:




Die Definition eines Ausschnitts erfolgt durch das Ziehen der Maus bei gedrückter linker Maustaste, wobei die Maus im Bereich der hier mit roten Pfeilen markierten Bereiche geführt werden muss.

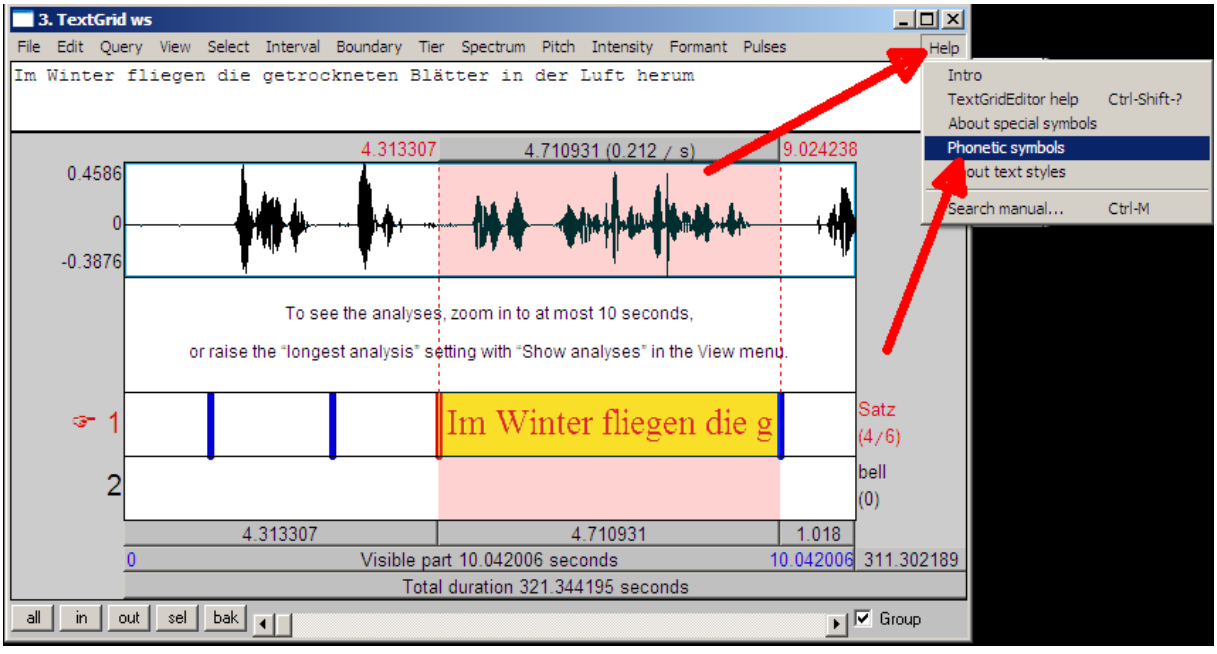
Um eine Intervallgrenze zu setzen, muss man mit der linken Maus wiederum in einen der beiden mit Pfeilen markierten Bereiche klicken. Es erscheint dann (außer einer horizontalen, für uns irrelevanten) eine gestrichelte vertikale rote Linie, die jeweils an ihrem Schnittpunkt mit den oberen Begrenzungslinien der Tiers (grüne Pfeile) einen Kreis aufweist (blaue Pfeile). Ein Klick in diesen Kreis erzeugt eine Intervallgrenze im entsprechenden Tier.

Die Intervallgrenzen werden durch vertikale Linien im entsprechenden Tier gekennzeichnet:

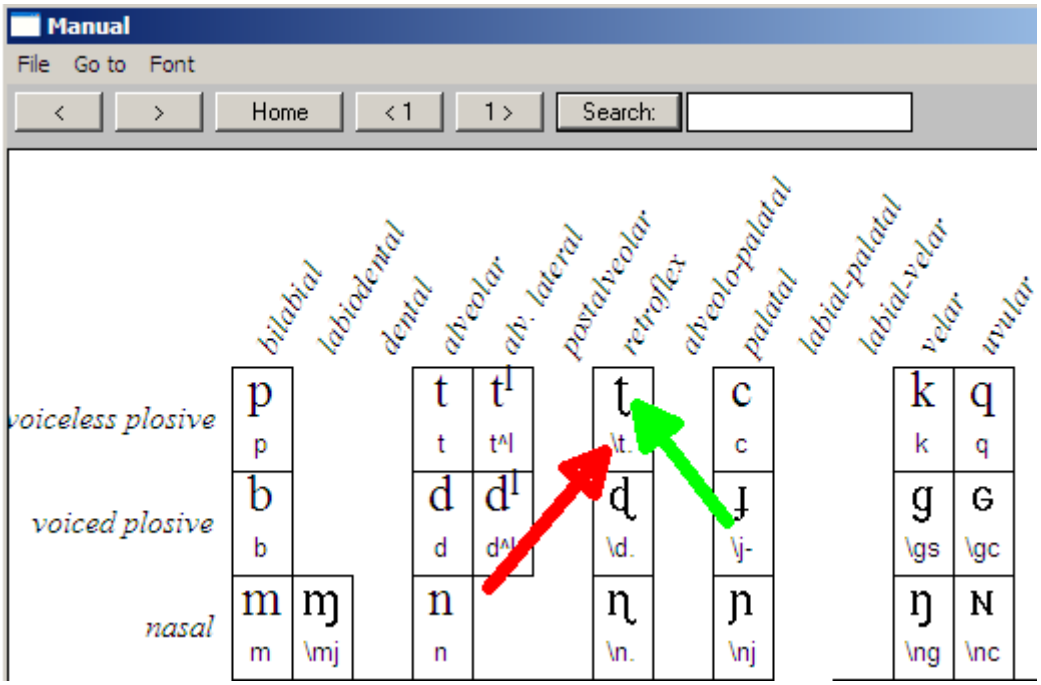



Die Auswahl eines Intervalls erfolgt durch Mausklick in den Bereich zwischen zwei vertikale Linien (= Intervallgrenzen). Anschließend erscheint das Intervall rot hinterlegt, und es ist möglich, in den weißen Bereich (roter Pfeil) am oberen Rand des Fensters den Transkriptionstext einzugeben. Schon bei der Texteingabe erscheint der Text zusätzlich im entsprechenden Bereich des Tiers (in roter Schrift).

Praat erlaubt es nicht, phonetische Sonderzeichen zu verwenden. Erforderlichenfalls müssen diese durch eine Abfolge von ASCII-Zeichen eingegeben werden. Die zu verwendenden Zeichentabellen können durch Klicken auf das Hilfe-Menü abgerufen werden:

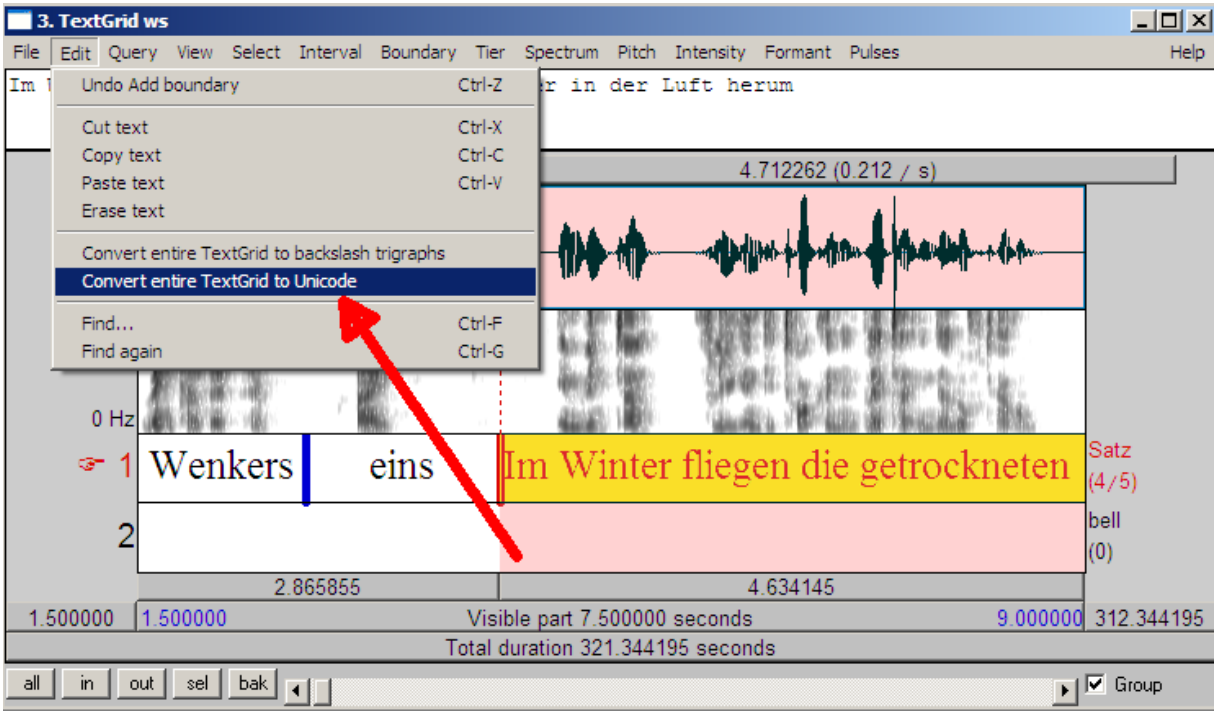


Hier ein Ausschnitt aus der Konsonanten-Tabelle:

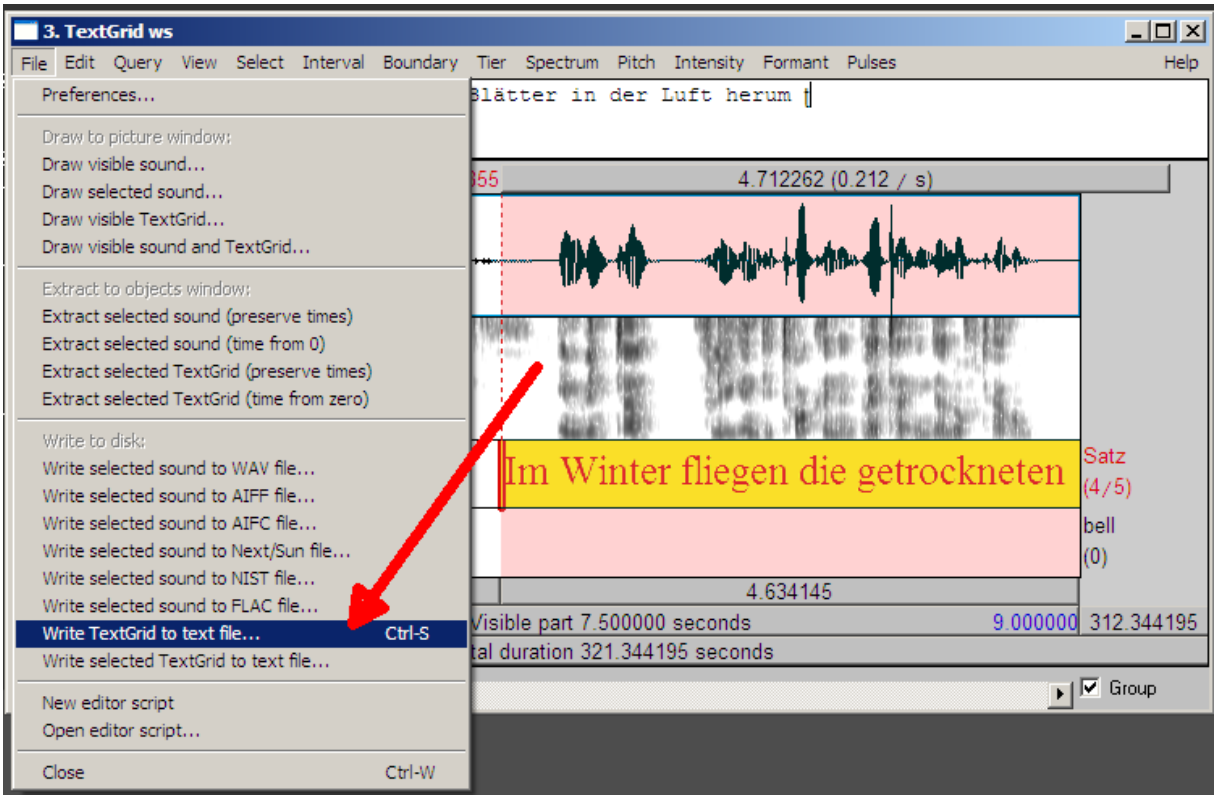



Der "retroflexe stimmlose Plosivlaut", dessen phonetisches Symbol hier mit dem grünen Pfeil markiert ist, muss in Praat mit der Zeichenfolge "\t." wiedergegeben werden.

Nach Abschluss der Transkriptionsarbeit kann der Text in eine sog. "TextGrid-Datei" exportiert werden. Es empfiehlt sich, den Text zuvor gemäß Unicode zu kodieren. Dieser Schritt erfolgt über das Menü "Edit" -> "Convert entire TextGrid to Unicode":




Anschließend wird die Transkription über das Menü "File" und "Save TextGrid as text file" in eine Textdatei geschrieben:



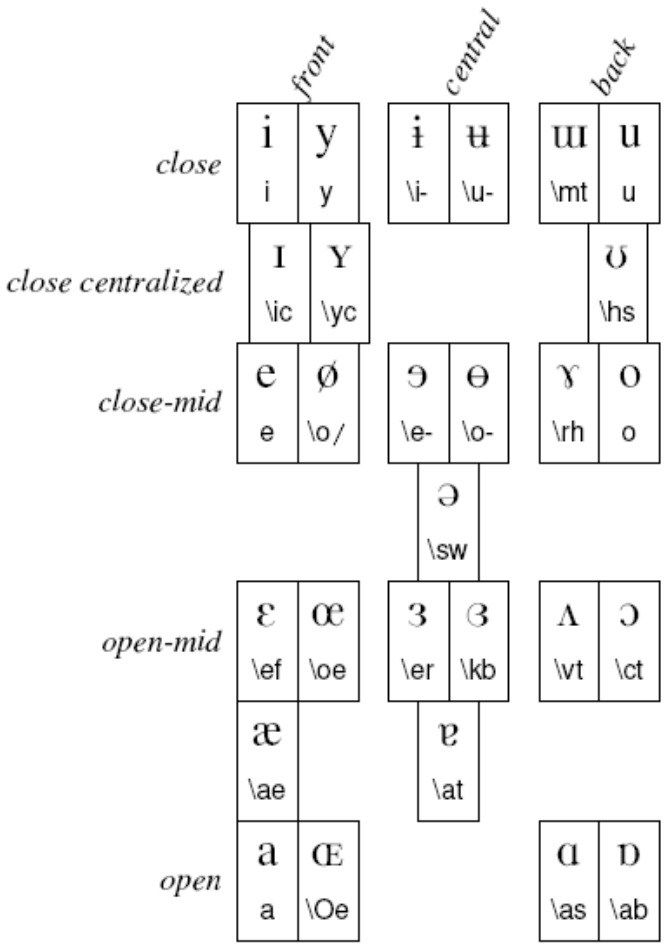

Die dabei entstandene Datei kann anschließend z.B. mit einem awk-Skript in Tabellenform gebracht und dann in eine Datenbank importiert werden:

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 321.3441950113379
tiers? <exists>
size = 2
item []:
    item [1]:
        class = "IntervalTier"
        name = "Satz"
        xmin = 0
        xmax = 321.3441950113379
        intervals: size = 5
        intervals [1]:
            xmin = 0
            xmax = 1.0630635245901638
            text = ""
        intervals [2]:
            xmin = 1.0630635245901638
            xmax = 2.8301618852459014
            text = "Wenkersätze"
        intervals [3]:
            xmin = 2.8301618852459014
            xmax = 4.365854508196721
            text = "eins"
        intervals [4]:
            xmin = 4.365854508196721
            xmax = 9.078116803278688
            text = "Im Winter fliegen die getrockneten Blätter in der Luft herum"
        intervals [5]:
            xmin = 9.078116803278688
            xmax = 321.3441950113379
            text = ""
    item [2]:
        class = "TextTier"
        name = "bell"
        xmin = 0
        xmax = 321.3441950113379
        points: size = 0
```

**Anhang II: Konventionen zur Erfassung phonetischer Sonderzeichen gemäß "Praat"**

1. Vokale



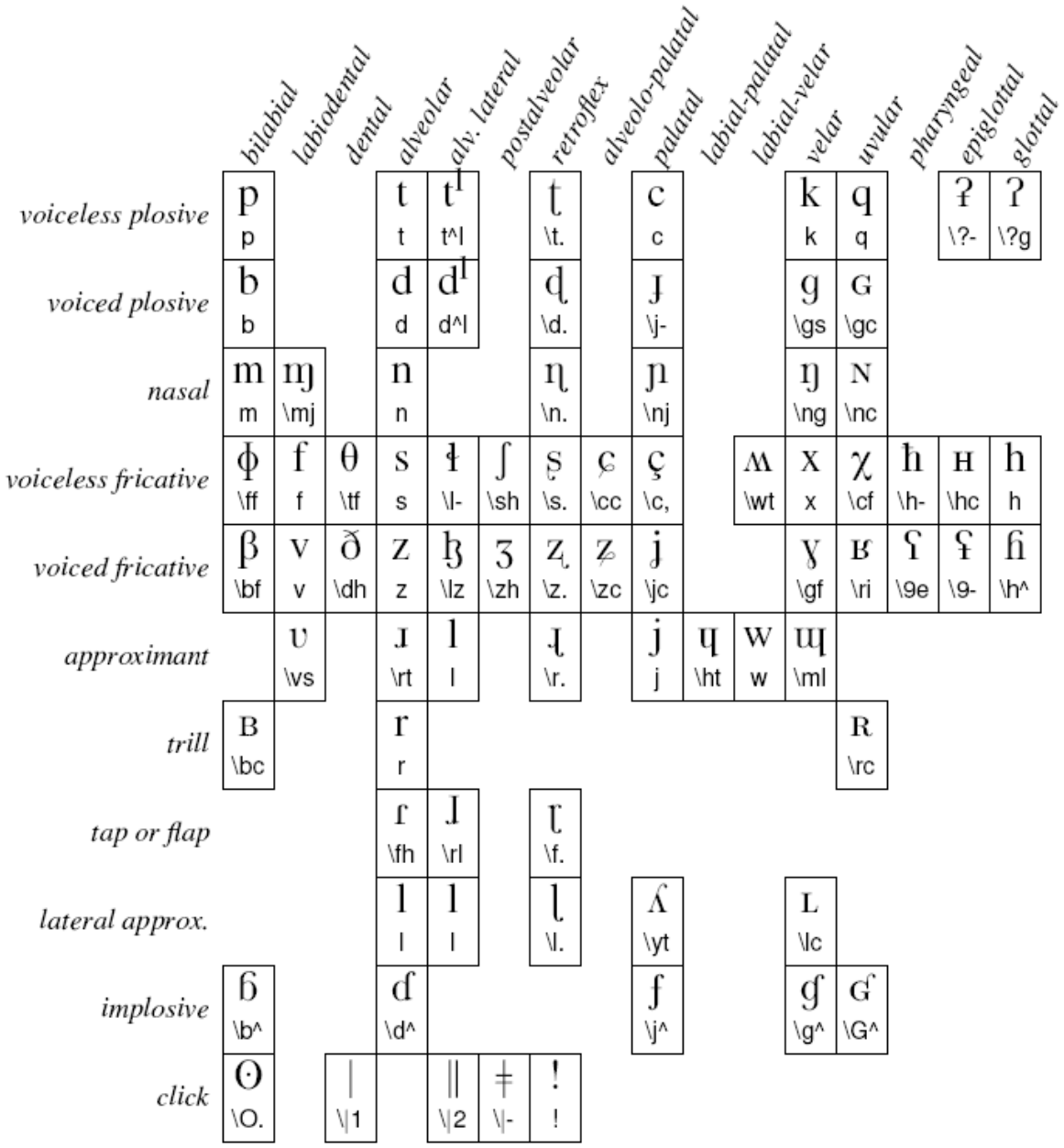

Other vowel symbols are:

ɚ \sr (*schwa with right hook*): rhotacized schwa

## How to remember the codes

For most of the codes, the first letter tells you the most similar letter of the English alphabet. The second letter can be *t* (*turned*), *c* (*capital*), *s* (*script*), *r* (*reversed*), *-* (*barred* or *retracted*), or */* (*slashed*). One symbol (ɛ) is a phonetic version of a Greek letter. The codes for ə, ɤ, ʊ and ɞ are abbreviations for *schwa*, *ram's horn*, *horseshoe*, and *kidney bean*.

2. Konsonanten




Other consonant symbols:

ɬ \l~ (*l with tilde*): velarized *l*

ɧ \hj (*heng with hooktop*): the Swedish rounded post-alveolar & velar fricative

## How to remember the codes

For most of the codes, the first letter tells you the most similar letter of the English alphabet. The second letter can be *t* (*turned*), *c* (*capital* or *curled*), *s* (*script*), - (*barred*), *l* (*with leg*), *i* (*inverted*), or *j* (*left tail*). Some phonetic symbols are similar to Greek letters but have special phonetic (*f*) versions with serifs (ɸ, β, ɣ) or are otherwise slightly different (θ, χ). The codes for ŋ (*engma*), ð (*eth*), ʃ (*esh*), and ʒ (*yogh*) are traditional alternative spellings. The retroflexes have a period in the second place, because an alternative traditional spelling is to write a dot under them. The code for ɾ is an abbreviation for fishhook.

3. Diacritica

In line:

ː \:f the phonetic length sign
ˈ \'1 primary stress
ˌ \'2 secondary stress
ǀ \|f the phonetic stroke
t̚ t\cn (*combining left angle above, corner*): unreleased plosive
ɜ˞ \er\hr (*combining rhotic hook*): rhotacized vowel

Understrikes:

n̩ n\|v (*combining vertical line below*): syllabic consonant
b̥ b\0v (*combining ring below*): voiceless (e.g. lenis voiceless plosive, voiceless nasal or approximant)
o̞ o\Tv (*combining down tack below, lowering*): lowered vowel; or turns a fricative into an approximant
o̝ o\T^ (*combining up tack below, raising*): raised vowel; or turns an approximant into a fricative
o̟ o\T( (*combining left tack below, atr*): advanced tongue root
o̠ o\T) (*combining right tack below, rtr*): retracted tongue root
e̱ e\-v (*combining macron below*): backed
o̟ o\+v (*combining plus sign below*): fronted
o̤ o\:v (*combining diaeresis below*): breathy voice
o̰ o\~v (*combining tilde below*): creaky voice
d̪ d\Nv (*combining bridge below*): dental (as opposed to alveolar)
d̼ d\Uv (*combining inverted bridge below*): apical
d̻ d\Dv (*combining square below*): laminal
u̯ u\nv (*combining inverted breve below*): nonsyllabic
e̹ e\3v (*combining right half ring below*): slightly rounded
u̜ u\cv (*combining left half ring below*): slightly unrounded

Overstrikes:

ɣ̊ \gf\0^ (*combining ring above*): voiceless
é \ep\'^ (*combining acute accent*): high tone
è \ep\‘^ (*combining grave accent*): low tone
ɛ̄ \ep\-^ (*combining macron*): mid tone (or so)
ɛ̃ \ep\~^ (*combining tilde*): nasalized
ɛ̌ \ep\v^ (*combining caron, haček, wedge*): rising tone
ɛ̂ \ep\^^ (*combining circumflex accent*): falling tone
ö o\:^ (*combining diaeresis*): centralized
ɛ̆ ɛ\N^ (*combining breve*): short
k͡p t͡s k\lip (*combining double inverted breve, ligature*): simultaneous articulation, or single segment