# Concordancing for Parallel Spoken Language Corpora

*Dafydd Gibbon, Thorsten Trippel*

Universität Bielefeld
gibbon@spectrum.uni-bielefeld.de
ttrippel@spectrum.uni-bielefeld.de

*Serge Sharoff*

Humboldt Fellow, Universität Bielefeld
Russian Res. Inst. for AI
sharof@aha.ru

## Abstract

Concordancing is one of the oldest corpus analysis tools, especially for written corpora. In NLP concordancing appears in training of speech-recognition system. Additionally, comparative studies of different languages result in parallel corpora. Concordancing for these corpora in a NLP context is a new approach. We propose to combine these fields of interest for a multi-purpose concordance for Spoken Language Data, opening the opportunity of combining corpus-linguistic and NLP methods resulting in a broader empirical basis for NLP research. Theoretic models for audio-concordances are discussed. Principles of the structure and design of a parallel audio concordance are given, coding by means of XML to ensure reusability and flexibility, using time stamps for referencing from annotations to the signal.

## 1. Introduction

One of the most commonly used corpus analysis tools, and certainly the oldest,[1] is the *text concordance*, traditionally defined as a table in which words which occur in a text are paired with citations of the text passages in which they occur. The art of concordancing has reached a peak in computational corpus linguistics, with access criteria which include not only word keys but also linear or hierarchical tagging, the choice of static (precompiled) concordances or dynamic (on-the-fly, free-key) concordances [1].

Users of speech corpora are not so fortunate, despite the fact that the field of *audio indexing* is developing rapidly, and despite concordance-like techniques for using annotated speech signals to access spoken language corpora in order to train stochastic models for automatic speech recognition. Most tagged spoken language corpora and treebanks still restrict themselves to transcriptions, i.e. textual representations, and do not in general provide access to the speech signal.

In this paper, we examine and define the notion of *audio concordance*, discuss an implementation, and suggest that a standardised approach to audio concordancing for unilingual and multilingual corpora would provide valuable heuristic support tools for a wide range of linguistic and information-retrieval activities. In particular, we address the question of concordances for parallel aligned speech corpora.

We use the German VERBMOBIL speech-to-speech translation corpus, a corpus with German, English and Japanese data, including both monolingual and multilingual dialogues. For testing we selected the dialogue *M872B* of [2], a bilingual-dialogue for English and German, 435 seconds (approx. 7 min.). The speech signals are transcribed (in Verbmobil terminology 'transliterated') and annotated following the VERBMOBIL conventions, which were converted to an XML format following [3].

For reusability in different contexts, XML based formats are used for the concordance, based on the formats specified in [4], extended for multinlingual and parallel texts.

[1]The technique dates back at least to the Middle Ages; the oldest reference to 'concordaunce' in this sense given by the Oxford English Dictionary is 1387.

## 2. Characterisations and definitions

### 2.1. Audio concordance

We start with some straightforward and fairly evident characterisations and move on to complex corpora and correspondingly more complex notions of concordance.

First, extending the traditional definition, we provide an initial (and partial) characterisation of audio concordance:

> An *audio concordance* is a table in which representations of units which occur in an annotated spoken language corpus are paired with citations from the annotations in which they occur.

For present purposes, our characterisation of *annotation* is:

> An *annotation* is a pair consisting of a *symbol* and a *time-stamp*, where the symbol represents some linguistic property of a speech signal and the time-stamp represents the temporal location of this property in the speech signal.

Thus, a transcription of a recording, and the recording which it transcribes, implicitly constitute a minimal annotation $< R, T >$, $R$ being a recording and $T$ being a transcription of it, if the start and end of the recording are understood to be aligned with the start and end of the transcription, respectively. In the general case, the time-stamps can be representations of temporal points (e.g. offsets from the beginning of a speech file) or intervals represented as pairs of points, and the properties can be any arbitrarily complex linguistic unit, from phones through syllables to words or longer units, or linear or hierarchical categories (i.e. generalisations over classes of such units). We refer to the hierarchy of annotation domains as *annotation granularity*.

Broadly speaking, in this characterisation we thus follow the *annotation graph* approach of Bird & Liberman [5] as applied to speech signals. We characterise annotation graphs as sets of annotations of the same recording.

We now turn to some details in preparation for a discussion of the concordancing of parallel corpora.

> Let $LEX^m$ be a set of symbolic representations (e.g. words) in a language $L^m$, $LEX_i^m \in LEX^m$, and $C^m$ an annotated corpus with elements $C_j^m$ in $L^m$, which is an annotation of some

$LEX_i^m$ such that
$C_j^m =< property(LEX_i^m), < t_{start}, t_{end} >>$

A purely corpus-derived lexicon $LEX_c^m$ is generated by a function

$$f_{lex} : C^m \rightarrow LEX^m$$

In the simplest case, this is obviously just a wordlist generating function. The function may be arbitrarily sophisticated in terms of morphological and collocational analysis, in connection with additional external information such as distributional analysis procedures, or lexical rules from lexical morphology, phonology or semantics. A concordance is then a kind of inverse function; more precisely $f_{conc} : L_c \rightarrow \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ is the power set of the contexts in $X$.

### 2.2. Audio concordance access

We need to define an *audio access function* which the audio concordance provides to the speech signal. For this we need a more detailed definition, following event logic conventions [6], in which

$$f_{aconc} : LEX_c^m \rightarrow f_{play}(C_j^m, R_j^m)$$

where $C_j^m =< property(LEX_i^m), < t_{start}, t_{end} >>$ and $R_j^m = f_{signal}(t_{min}, t_{max})$ is a recording such that $t_{min} \leq t_{start} \wedge t_{end} \leq t_{max}$ (i.e. the interval aligned with $C_j^m$ is fully contained in the interval aligned with $R_j^m$).

As yet there are no widely available tools corresponding to this specification; this is surprising, since the conceptual basis for the specification is rather straightforward, and in fact *ad hoc* tools roughly matching the specification are to be found in any speech recognition or synthesis development lab.

### 2.3. Multilingual audio concordance

Multilingual concordances serve as platforms for comparative studies of languages. Earlier this only included the possibility of comparing tagged data, coded in texts. The disadvantage especially for natural languages is that tagging is only possible according to state-of-the-art categories and more often only subsets of these categories were used according to specific requirements in a given research context. Providing a way to access the signals from the tags would provide a new kind of reusability in multiple dimensions: signals can be reused not from scratch and previous analyses can be linked and used for reference. For example morphophonological information for each language can be added to a word-level analysis by using existing information previously gained.

Text alignment of parallel texts combines NLP problems with issues such as signal recording, storage formats, annotation, etc. The combination of both fields results in an even more complex set of problems. On the other hand, there exist established techniques which can be deployed ([7]).

### 2.4. Possible users for an audio concordance

Audio concordances may serve in training phoneticians (availability of tagged corpora for evaluation of personal results), language learners (in hearing authentic material pronounced by native speakers with authentic intonation in context), language trainers (teaching materials), corpus linguists (doing analysis with corpora), applications for the blind (with the ability of merging a lexicon with a text-to-speech system and examples from an audio concordance) and for general reference purposes.

Parallel audio texts may serve for comparative studies, training in translation, various machine translation methods, language learners (in seeing two languages at a time), and people interested in languages in general.

More generally a multilingual audio concordance enables the researcher to search for audio patterns in a parallel corpus, as long as the corpus is tagged sufficiently. Therefore theoretical models and theories can be verified empirically independently of singularities or oddities in limited examples.

We proceed by characterising parallel corpora, and developing a definition of multilingual audio concordance from this characterisation.

## 3. Parallel corpora for spoken language

In the case of written language, parallel corpora (corpora containing text pairs which are translation equivalents) are ubiquitous: literary, religious, scientific and commercial documents provide obvious sources of parallel texts, as there are strong commercial, scientific and cultural interests in the accurate translation of many types of written text. In the case of spoken language, parallel corpora are much less common because most spoken discourse is intended to be physically transient rather than persistent, and the interest in keeping a corpus for these kinds of discourse is limited for example to cases of learning a foreign language, when the education practices consist in presenting recorded speech to learners together with its written translation into the native language of learners or to spoken language system development..

In the case of interpreted speech, typically in formal institutional contexts, this situation changes, and the well-known case of the Canadian Hansard texts are in principle a prime case of a parallel spoken language corpus - except that the texts are not aligned with speech recordings. So the question of what constitutes a parallel spoken language corpus is not obvious.

One type parallel spoken language corpus occurs in a scenario where two different language-speakers try to communicate with each other through an interpreter, who provides direct speech-to-speech translation from one language into the other. But there are many differences between interpreted translation and written translation, some of which are due to domain differences between written and spoken language (e.g. its dialogue character, the presence of prosody and paralinguistic features, discourse marking items, edited and non-edited disfluencies), and others of which are due to temporal constraints (e.g. the lack of time for translation planning, consultation drafting and editing); see [7], p. 81f. Although there is a common core of lexico-syntactic translation equivalence problems with written and spoken language, speech-to-speech translation may well be much less accurate and error-prone than written language translation for the reasons given, and specifically, there may be considerable information loss due to non-translation or mistranslation of prosody, paralinguistic features, discourse markers, etc., in addition to the usual sources of non-translation or mis-translation.

For the purpose of concordancing, we understand the notion of *unit* (cf. the initial characterisation above) to be any property of speech, whether prosody, paralinguistic property, phrase, word, syllable, etc. The archetypal unit for a concordance is the (simplex, derived or compound) word, but although our discussion focusses on the word, the validity of the discussion is more general, and may include phrases, prosody or discourse markers..

We take parallel corpus concordancing to depend crucially

on the definition of a *consistent vocabulary* in the technical sense of the term [3], i.e. a vocabulary which is first, *corpus-based*, second, *controlled for content*, and third *formally well-defined*. The definition of *multilingual consistent vocabulary*, also given in [3], is more complex. First, the corpus is a parallel corpus. Second, control for content involves the notion of translation equivalence. Third, formal definition involves a translation function.

The approach that was taken for the Verbmobil lexicon is described in [8]. A *translation equivalent of a given wordlist* from the source language is the list of words needed for translation into a target language. *Equivalence* in this sense does not mean a one-to-one relation between words but more general a relation between lists of words in texts constituting a parallel corpus.

We come to the definition of *parallel corpus*. A parallel corpus for languages is a corpus $< L_k^i, L_l^j >$ where for some symmetrical mutual translation (translation equivalence) function $f_{mtr} : L^i \rightarrow L^j$ the following assertion holds:

$$\forall L_k^i \in L^i, \forall L_l^j \in L^j (f_{mtr}(L_k^i) = L_l^j \wedge f_{mtr}(L_l^j) = L_k^i)$$

We do not assume a literal compositionality function for translation in which the same condition holds for the vocabulary, in which the translation equivalence function for corpora is a function of the translation equivalence for words. Rather, we use the weaker definition of a *multilingual consistent vocabulary*, in which $LEX_i^m$ and $LEX_j^n$ constitute a multilingual consistent vocabulary if and only if

$$f_{mtr}(L_k^i) = L_l^j) \wedge f_{mtr}(L_l^j) = L_k^i) \text{ and}$$
$$LEX_i^m = f_{lex}(C_j^m) \wedge LEX_k^n = f_{lex}(C_l^n)$$

This is a somewhat extreme definition, in which the entire corpus is the domain for translation equivalence, and the notion of lexical unit can therefore be quite different for the two corpora (a useful feature where typologically very different languages are involved). Realistically, though, the granularity for translation will be somewhere between the entire corpus and the lexical unit, for instance the utterance (in a spoken corpus), or perhaps the adjacency pair or exchange (for ritual dialogue constituents). To simplify discussion, we will not distinguish between a whole corpus which is translationally equivalent and a component of a corpus which is the maximum domain of translation equivalence. For more information on translation equivalence see [8]. We refer to the relation between different domains of annotation, from the entire corpus down to the smallest lexical unit, as the *translation equivalence granularity*.

# 4. A parallel corpus concordance

## 4.1. Concordance design as document design

A concordance is a formally specified and (semi-) automatically generated *document* designed for human or machine reading. In this respect, it differs fundamentally from *ad hoc* audio indexing tools with application specific formats. Basic principles of document design distinguish between (at least) three levels of organisation for such documents (Figure 1): *Content Structure* (CS), i.e. the semantic network, database, etc. which characterises the document content; *Documentation Structure* (DS), i.e. the underlying hierarchical and heterarchical document architecture; *Presentation Structure* (PS), i.e. layout and rendering; cf. [9]. The underlying *DS* is mapped into the content model *CS* (*c-semantics*) and the presentation model *PS* (*p-semantics*). The *CS* of a parallel corpus concordance has been

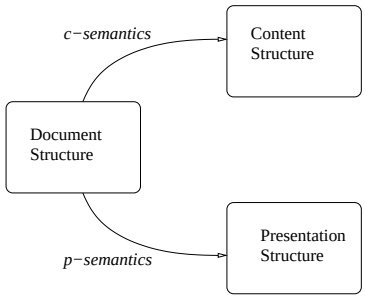


Figure 1: *Basic document design components.*

sufficiently characterised above for present purposes. In designing a parallel corpus concordance, we discuss a number of preliminary issues and then define the *DS* and give an example of a *PS* implementation.

## 4.2. Structure of the concordance

The Document Structure of a parallel concordance is determined by a number of factors, which have been introduced above.

1. two signal sources;
2. translationally equivalent annotations of the signal sources, constructed in accordance with
   - the annotation granularity level of the domain of maximum translation equivalent;
   - the annotation granularity level of lexical unit (e.g. the word);
3. temporal annotation conventions [6] in terms of a partial ordering in relative time (i.e. in terms of temporal precedence and overlap) or in terms of absolute time (i.e. time stamps).

From the structural point of view we can distinguish two different levels of abstraction in the annotation of a parallel corpus:

1. the monolingual corpus representation level where the signal information, transcription and categorial tag information are connected to each other and
2. the parallel text level where the separate languages are referenced to each other.

This division is necessary because annotation alignment is based on two different principles of temporal organisation [6]: the first level is annotated in terms of absolute time on a time stamp basis, while alignment on the second level results from relations between successions of maximal domains of translation equivalence, for instance where a turn is followed by its translation, which is followed by a response, etc., or, in simultanous interpreting, where a turn is accompanied by its slightly delayed translation, etc.

## 4.3. Content structure

A concordance which stores a written corpus implies a mapping from (orthographic) forms to annotations of their linguistic properties, like part-of-speech tags, grammatical functions or lexical meanings. For morphological properties, the approximation is in general precise (though a part-of-speech tagger for English may fail to distinguish between uses of *left* as an adverb or as a past participle, for example). For grammatical and semantic properties, data loss in the process of annotation is more frequent. In addition to this mapping from orthographic

forms to linguistic properties, a concordance for spoken language implies another layer: sound forms are mapped into their transcriptions as orthographic forms. On the one hand, all transcriptions result in loss of information, but access to the acoustic signal stored in the concordance potentially compensates for the loss and additional information can still be extracted from the signal as required. This means that initial annotation work can be quite economical because adding information on a later stage is still possible by reaccessing the signal. A simple proof-of-concept implementation in `Perl`, based on the present specification, can be found in [4] with sentence level alignment. Re-annotation becomes easier by using (and enhancing) existing annotations (for automatic enrichment of information see [3]).

The transcribed information can be used for adding extra information based on computational morphology, lexicography and statistical methods. By using this technology corpus linguistic approaches become available to signal processing.

## 5. The case of discourse particles

The transcription itself is based on requirements of the proper phonemic encoding and may in fact be orthographic for many purposes. Additional levels of annotation granularity are more task dependent. One example of annotation requirements which depend on task and speech style is given by discourse particles, like segmentation markers or interjections (the terminology follows [10]). The discourse particles are particularly important for NLP because they are common in natural communication, significant problems for speech recognition and later syntactic processing. At the same time they are not covered by traditional grammars, which are primarily oriented towards written texts. A speech concordance provides a possibility to study their use empirically.

The encoding of German interjections used in the Verbmobil corpus is based on their phonetic properties, for example:

```
<"ah> - vocalic articulation,
        independent of the vowel quality,
<"ahm> - vocalic articulation +
         nasal articulation.
```

The above mentioned interjections are interpreted in English typically as *well* or *yeah* and are encoded without any concern for details of their phonetic properties. At the same time, the encoding of meanings for such interjections are based on their function in the dialogue communication as pragmatic markers. In terms of its linguistic properties, *well* in English or *äh, ähm* in German can function as, for example:

- FI - filling marker, which indicates that the speaker proceeds with the next statement and wishes to continue the ongoing discourse.
- FR - frame marker, which is used to initiate a narrative segment.
- RF - reformulation marker, which indicates a reformulation of an earlier statement.

Compare the function of *well* in the two sentences from [2]:

(1)*okay. well, it doesn't sound that bad.*

(2) *well, are there any special deals?*

In the first case, *well* indicates a pause in the speech, when the speaker considers whether to accept the offer [FI], while in the second case, it is aimed at the initiation of a new topic for discussion [FR].

It is unlilkely that more than a very minimal automatic annotation of discourse particles in terms of their discourse functions is feasible. As the study of interjections in the Verbmobil corpus [10] shows, *ja* in German can fulfill all identified functions depending on the context and the intentions of the author. At the same time, there are some fairly reliable indications that help to suggest a possible function of a marker. For instance, *äh* occurring within a word with a new syntactic group following it clearly functions as a repair marker:

(3) *tja, welcher Tag käme dann wohl für Sie in die Fra<äh> in Frage für das Seminar? (well, which day would then be under the co/under consideration for you for the seminar?)*

On the other hand, when *äh* occurs turn-initially, it typically functions a frame marker, as in the translation of (2):

(4) *<äh> gibt es zufällig irgendwelche besonderen Angebote?*

## 6. Conclusion

The fact that a multilingual speech concordance relates the signal, its transcription and linguistic properties provides a method for accessing the signal file from the lexicon via the annotation: a lexicon-entry or tag is used as a lookup in the annotation, all occurences of the lookup are found and the results are displayed from the annotation in relation to the signal. By using this technique it is possible to relate parallel spoken language fragments for many purposes.

## 7. References


[1] Dafydd Gibbon, "Computational lexicography," in *Lexicon Development for speech and language processing*, Frank Van Eynde and Dafydd Gibbon, Eds. Kluwer Academic Publishers, 2000.

[2] "Verbmobil CD 32.0," CDROM, June 1999.

[3] Dafydd Gibbon, Harald Lüngen, and Andreas Witt, "Enhancing speech corpus resources with multiple lexical tag layers," in *Proceedings of LREC 2000*, Athens, 2000.

[4] Thorsten Trippel, Nils Jahn, Dafydd Gibbon, and Soma Ouattara, *Preliminary Specification, Proof–of–Concept Implementation of a Portable Audio Concordance (PAC)*, Universität Bielefeld, Bielefeld, 2001, http://coral.lili.uni-bielefeld.de/langdoc/SPECIFICATION/.

[5] Steven Bird and Mark Liberman, "Towards a formal framework for linguistic annotations," in *Proceedings of the 5th International Conference on Spoken Language Processing*, 1998.

[6] Julie Carson-Berndsen, *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*, Kluwer Academic Publishers, Dordrecht, 1998.

[7] Dafydd Gibbon, Roger Moore, and Richard Winski, Eds., *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, 1997.

[8] Dafydd Gibbon and Harald Lüngen, "Speech lexica and consistent multilingual vocabularies," in *Verbmobil: Foundations of Speech-to-Speech Translation*, Wolfgang Wahlster, Ed. Springer, Berlin, Heidelberg, 2000.

[9] Dafydd Gibbon, Harald Lüngen, and Andreas Witt, "Enhancing speech corpus resources with multiple lexical tag layers," in *Proceedings of LREC 2000*, Athens, 2000.

[10] Kerstin Fischer, *From Cognitive Semantics to Lexical Pragmatics: functional polysemy of discourse particles*, Mouton de Gruyter, Berlin, New York:, 2000.