

Intonationsstrukturen
in
Sprachdialogsystemen

von
Zeno Leander Wolze
(2007)

Einleitung

Bei der Gestaltung eines Sprachdialogsystems steht der Designer vor der Herausforderung, die einzelnen Pfade, die durch das System führen, sprachlich möglichst flüssig zu gestalten. Eine solche flüssige Gestaltung ist nicht nur abhängig von Dialog-Fluss und Text, sondern auch – und dies wird meistens außer Acht gelassen – von der Intonationsstruktur. Der vorliegende Aufsatz beschreibt, welche Funktionen der Intonationsstruktur im Gesprächsdiskurs zukommen und worauf beim Design und bei der Produktion von Voice-Prompts zu achten ist.

Flüssige Gestaltung – eine Definition

Um den Beitrag der Intonationsstruktur zu einer flüssigen Gestaltung von Dialogen zu beschreiben, muss zunächst einmal erläutert werden, was generell unter „flüssige Gestaltung“ oder genauer „flüssige Textgestaltung“ verstanden werden soll. Gemeint ist, dass der Kontext (vorangeangene und nachfolgende Dialog-Zustände) in angemessener Weise in die Gestaltung der einzelnen Textpassagen mit einbezogen wird. Ein viel beschriebenes sprachliches Mittel, um dies ein Stückweit zu gewährleisten, sind die sogenannten Diskursmarker¹. Diese Textelemente signalisieren zum Beispiel, dass eine unmittelbar zuvor vom User ausgeführte Aktion vom System registriert wurde und leiten gleichzeitig zum nächsten Dialogschritt über:

- (1) System Bitte nennen Sie Ihre Kontonummer.
User 345243568
System *Vielen Dank*. Bitte nennen Sie *nun* Ihre Pin.

Ohne die Diskursmarker erscheinen die Dialog-Schritte viel stärker wie eine monotone Aneinanderreihung von Abfrageschritten. Das System wirkt maschinenhafter, wirkt so, als habe es kein „Gedächtnis“, weil es mit keinem Wort Bezug nimmt auf den unmittelbar vorangegangenen Gesprächsverlauf:

- (2) System Bitte nennen Sie Ihre Kontonummer.
User 345243568
System Bitte nennen Sie Ihre Pin.

¹ Diskursmarker sind sprachliche Ausdrücke, deren Vorkommen der Organisation, Gliederung, Strukturierung vor allem gesprochener Sprache dient (vgl. Metzler, Lexikon der Sprache)

Flüssige Textgestaltung entsteht also durch adäquaten Einbezug des Kontextes in eine einzelne Textpassage. Aber nicht nur beim Gestalten der Prompt-Texte muss der Kontext mit einbezogen werden, sondern auch bei der Intonierung derselben. Und genau dies stellt in der Praxis häufig ein Problem dar:

Wenn die Texte – nach Ihrer Fertigstellung durch den Dialog-Designer – in Produktion gehen, werden sie dem Sprecher und dem Aufnahmeleiter in der Regel als einzelne Fragmente ohne jegliche Kontextinformation vorgelegt. Eine Einweisung in die Dialog-Zusammenhänge würde Zeit in Anspruch nehmen und die Kosten für das Recording sollen ja so gering wie möglich gehalten werden. Folge kann ein Dialog sein, der an bestimmten Stellen stark unrund klingt, obwohl Dialog-Fluss und Text gut durchdacht sind.

Merkmale von Intonationsstrukturen

Unter der Intonation einer sprachlichen Äußerung wird allgemein der Verlauf der Grundfrequenz (auch: „Tonhöhe“), der Verlauf der Lautstärke und die zeitliche Dauer verstanden. Die Merkmale von Intonationsstrukturen lassen sich tendenziell zwei Klassen zuweisen:

In die erste Klasse fallen diejenigen Merkmale, deren Funktion primär linguistisch-semantic ist. Man spricht in diesem Falle auch von Betonungsstruktur. In die zweite Klasse fallen diejenigen Merkmale, die eher paralinguistische Funktion haben. Diese Merkmale beziehen sich auf den metakommunikativen Teil der Sprache. Sie sagen etwas aus über die Beziehung zwischen Sprecher und Hörer oder über den Sprecher selbst. Beide Arten von Intonationsstrukturen und ihre Funktionen, speziell im Rahmen von Sprachdialogsystemen, werden im Folgenden vorgestellt.

Betonung

Rein signalphonetisch versteht man unter „Betonung“ das akustische Hervorheben eines Ausschnitts einer Äußerung. Ein solches Hervorheben zeichnet sich aus durch einen Anstieg der Grundfrequenz, einen Anstieg der Lautstärke und einen Anstieg der zeitlichen Dauer (Dehnung) zugleich.

Betonungsmuster sind so selbstverständlich in unserem intuitiven, sprachlichen Wissen verankert, dass vor allem die untrainierten Sprecher des Deutschen (d.h. Sprecher ohne entsprechendes sprachliches Wissen) häufig Probleme haben, objektiv den Betonungsverlauf eines Wortes oder eines Satzes zu bestimmen. In manchen Fällen kann der genaue Betonungsverlauf sogar in einer akustischen Analyse nicht eindeutig ausgemacht werden. Aus diesem Grunde sind in den nachfolgenden Beispielen auch nicht immer die exakten Abschnitte angegeben, die am stärksten durch die Betonung

hervorgehoben werden (Hauptbetonung), sondern lediglich die Region um diese jeweiligen Abschnitte herum.

Es können zwei Arten von Betonungsverläufen unterschieden werden: Die Wortbetonung und die Satzbetonung. Die Wortbetonung ist relativ strikt festgelegt und daher im Rahmen der hier angestellten Betrachtungen wenig interessant.

Für die Satzbetonung gilt folgende Faustregel: Die Hauptbetonung eines Satzes liegt immer auf dem Teil des Satzes, der neue Information in den Gesprächsdiskurs einführt. Man spricht in diesem Zusammenhang auch von „Satzfokus“. Da neue Information im Deutschen in der Regel am Satzende zu finden ist, liegt dort normalerweise auch die Hauptbetonung des Satzes:

(3) Vielen Dank. Bitte nennen Sie nun Ihre Pin.

In diesem Beispiel stellt der Ausdruck „Pin“ zwar nicht die gesamte neue Information, wohl aber den Kern der neuen Information dar, weshalb ihm die Hauptbetonung des Satzes zukommt.

Nun ist es aber möglich, dass der Satzfokus nicht mit dem Ende des Satzes zusammenfällt, wodurch sich auch die Satzbetonung verschiebt:

- (4) (a) Ich habe den Hund gesehen (und nicht Du).
(b) Ich habe den Hund gesehen (und nicht die Katze).

In Beispiel 4a könnte der Gesprächspartner zuvor etwa gesagt haben: „Ich war derjenige, der den Hund gesehen hat.“. In Beispiel 4b hingegen könnte der Gesprächspartner zuvor gesagt haben: „Ich hab’ gehört, Du hast die Katze gesehen.“. Jeder Variante liegt also ein völlig anderer Gesprächskontext zugrunde, der bekannt sein muss, um diesen Sätzen die richtige Intonation zu verleihen.

Die Funktion der Satzbetonung ist es also, die Bedeutung zu verfeinern, indem sie die Basis-Bedeutung des Satzes in den größeren Zusammenhang des gesamten Gesprächsdiskurses stellt. Aus diesem Grunde wird die Satzbetonung auch im Rahmen von Sprachdialogsystemen relevant.

- (5) System Herzlich Willkommen beim Weckservice. Möchten Sie einen Weckauftrag aufgeben oder einen Weckauftrag löschen?
- User Einen Weckauftrag aufgeben.
- System Gut. Wann genau möchten Sie denn geweckt werden?
(falsch)
- System Gut. Wann genau möchten Sie denn geweckt werden?
(richtig)

In der als falsch markierten Variante liegt die Betonung (wie normalerweise üblich) am Satzende. Dies passt jedoch nicht in den gegebenen Gesprächskontext: Dass der User geweckt werden möchte, ist schon bekannt, ist also alte Information. Die Betonung muss daher im ersten Teil des Satzes liegen, in dem es um den genauen Zeitpunkt des Weckens geht.

In diesem Beispiel hat die falsch gesetzte Betonung vermutlich einen vergleichsweise geringen Effekt. In den meisten Fällen dürfte dem User nur unterbewusst auffallen, dass das Gespräch hier nicht richtig „rund“ läuft, dass also das Diskurswissen (das Wissen über das bis hierhin stattgefundene Gespräch) des Systems nur simuliert ist. Eine falsche Satzbetonung kann jedoch auch stärker ins Gewicht fallen:

- (6) System Von wo nach wo möchten Sie fahren?
- User Von Berlin Zoo nach äh – na – Wolfsburg Hauptbahnhof
- System Entschuldigung, bitte sagen Sie mir noch einmal von wo nach wo Sie fahren möchten.
- User (brüllt) Von Berlin Zoo nach Wolfsburg Hauptbahnhof!
- System Tut mir leid, ich kann Sie im Moment wirklich schlecht verstehen. Am besten, wir gehen das Ganze Schritt für Schritt durch. Bitte nennen Sie als erstes die Stadt oder Gemeinde, in der Sie losfahren möchten.
(falsch)
- System Tut mir leid, ich kann Sie im Moment wirklich schlecht verstehen. Am besten, wir gehen das Ganze Schritt für Schritt durch. Bitte nennen Sie als erstes die Stadt oder Gemeinde, in der Sie losfahren möchten.
(richtig)

Durch die Phrase „Schritt für Schritt“ wird signalisiert, dass das System in einen anderen Gesprächsmodus schaltet. Statt eine (relativ) offene Frage zu stellen, werden die Informationen nun stückweise in einem geführten Dialog erfragt. Die Phrase „Schritt für Schritt“ ist demnach also der Kern der neuen Information und muss daher in diesem Satz herausgehoben werden. Die Betonungsstruktur in der falschen Variante hat folgenden Effekt: Der Fokus liegt auf dem Verb „durchgehen“, wodurch die Information, dass überhaupt eine Abfrage stattfinden soll, als neu markiert ist. Dies ist deswegen falsch, weil im vorangegangenen Diskurs ja bereits zweimal der Versuch unternommen wurde, eine Abfrage durchzuführen. Als Nebeneffekt werden dadurch auch die Bemühungen des Users,

eine gültige Eingabe zu tätigen, nicht gewürdigt. Dies fällt vor dem Hintergrund, dass dieser ohnehin durch den bisher nicht gelungenen Dialog verstimmt ist, relativ stark ins Gewicht.

Nicht nur Dialog-übergreifend, sondern auch innerhalb einer Textpassage bzw. eines Prompts hat die Betonungsstruktur eine wichtige Funktion. Gängiges Beispiel ist hier das akustische Hervorheben von Keywords, was „automatisch“ geschieht, wenn die Keywords wie üblich am Satzende stehen:

- (7) Wenn Sie einen Weckauftrag aufgeben möchten, dann sagen Sie „aufgeben“. Möchten Sie einen Weckauftrag löschen, dann sagen Sie bitte „löschen“.

Zusätzlich zur Betonung kann der Sprecher das Keyword auch noch durch eine kleine Pause vom Rest des Textes abtrennen. Man spricht in diesem Zusammenhang auch von „akustischen Anführungszeichen“. Die Satzendstellung der Keywords hat hier, nebenbei gesagt, noch zwei weitere Effekte: Die neue Information (hier: die Keywords) ist dadurch, dass sie zuletzt genannt wird, immer am frischesten im Gedächtnis. Außerdem kann sich der User so zuerst die mit dem Keyword verbundene Option anhören, bevor er sich entscheiden muss, ob er sich das Keyword merken möchte oder nicht. Beides sind bekannte Prinzipien des VUI-Designs.

Wenn die Keywords im Text nicht so eindeutig auszumachen sind wie in Beispiel 7, muss der Sprecher wieder ein Stückweit Kontextwissen besitzen, um die Keywords als solche zu erkennen und entsprechend zu betonen:

- (8) Herzlich willkommen beim Weckservice. Möchten Sie einen Weckauftrag „aufgeben“ oder einen Weckauftrag „löschen“?

Werden die Wörter „aufgeben“ und „löschen“ beide betont, ist klar, dass sie die Keywords darstellen und dass der User die Wahl zwischen entweder der einen oder der anderen Option hat. Bei einer falschen Betonung wird diese Frage zur ja/nein-Frage. Ein User, der entweder das eine oder das andere will, wird mit „ja“ antworten, wodurch – bei einer entsprechenden Gestaltung der Grammatik – der Dialog nicht gelingen kann.

Anders als in den obigen Beispielen ist hier nicht vergangenes, sondern zukünftiges Wissen gefragt. Der Sprecher muss wissen, welche Optionen gewählt werden können, „ja“ und „nein“ oder aber „aufgeben“ und „löschen“.

Aber nicht nur im Zusammenhang mit Keywords und im größeren Diskurs spielt die Betonung innerhalb einer Textpassage eine Rolle. Bei inhaltlich komplizierten Texten trägt

sie dazu bei, dass der Hörer den Text besser strukturieren und dadurch besser verstehen kann:

- (9) Ein Gespräch kostet Sie als Anrufer aus dem T-Com Festnetz für Gespräche ins Festnetz 49 Cent für die ersten fünf Minuten, anschließend 4,9 Cent je angefangene Minute. Für Gespräche in die Mobilfunknetze zahlen Sie 49 Cent je angefangene Minute.

Dieser Text teilt komplexe Sachverhalte, die für die meisten User nicht alltäglich sein dürften, durch relativ wenige Worte mit. Es liegt also dicht gedrängte Information über ein nicht vertrautes Thema vor. Durch eine adäquate Betonungsstruktur wird es dem User erleichtert, den Text erstens in besser verarbeitbare Sinneinheiten zu zerlegen und zweitens diese einzelnen Einheiten zueinander in Beziehung zu setzen, sich also kurz gesagt die Struktur des Textes zu erschließen. Dies soll anhand des Textes in Beispiel 9 kurz illustriert werden: Durch die Betonung der Präposition „ins“ zum Beispiel wird der Beginn einer neuen Sinneinheit signalisiert und, das steckt in der Semantik von „ins“, es wird signalisiert, worum es in dieser Sinneinheit geht – nämlich um den „Ort“, in den das Gespräch aus dem Tcom-Festnetz gelangt (das Festnetz). Durch die Betonung von „ersten“ und „angefangene“ werden ebenfalls zwei Sinneinheiten herausgehoben, nämlich diejenigen, in denen es um die zeitlichen Bedingungen für verschiedene Arten der Abrechnung geht. Durch ihre Betonung werden diese Sinneinheiten aber nicht nur hervorgehoben, sondern gleichzeitig auch thematisch gegenübergestellt. Analog verhält es sich mit der Betonung von „Mobil“: Diese Sinneinheit wird „Festnetz“ und somit dem gesamten bisherigen Text gegenübergestellt.

Die Betonungsstruktur spiegelt also die semantische Struktur des Textes wieder. Ohne eine richtige Betonung „rauscht“ ein solcher Text einfach am Hörer vorbei und wird vermutlich aufgrund seines trockenen Inhalts als lästig empfunden oder, schlimmer noch, es wird unterstellt, es handele sich hier um das „Kleingedruckte“ zu Kosteninformationen, das einem unbemerkt untergeschoben werden soll.

Also, nicht nur dann, wenn die Betonungsstruktur vom Dialog-Verlauf abhängt, sondern auch, wenn der Text innerhalb einer Textpassage bzw. eines Prompts sozusagen für sich steht, muss der Sprecher den Sachverhalt im Grunde erst verstehen, um die richtige Betonungsstruktur zu finden. Um komplizierte und zeitraubende Erklärungen zu vermeiden, macht es Sinn, dass der Designer der Anwendung während des Recordings anwesend ist. Ist dies nicht möglich, sollten Regieanweisungen formuliert werden, indem betonte Elemente einfach im Text hervorgehoben werden. Sollte sich herausstellen, dass der Sprecher mit einer solchen Darstellung Probleme hat, kann die markierte Variante des Textes auch nur dem Aufnahmeleiter zu Verfügung gestellt werden. Allerdings muss der Aufnahmeleiter

zuvor informiert werden, was unter Satzbetonung akustisch zu verstehen ist: Die meisten Sprecher und Aufnahmeleiter haben eine eher „naive“ Auffassung von Betonung, so dass zumeist davon ausgegangen wird, dass die jeweiligen Passagen unnatürlich deutlich betont werden sollen.

Bei Betonungsstrukturen, die nicht vom Gesprächsverlauf abhängen (wie in Beispiel 9), macht es eher Sinn, den Aufnahmeleiter und den Sprecher dazu anzuhalten, den Text vor der Aufnahme erst einmal durchzulesen und zu verstehen. Hier würde ein Text-Markup aufgrund seiner Komplexität vermutlich nur zu Verwirrungen führen.

Paralinguistische Intonationsmerkmale

Die paralinguistischen Intonationsmerkmale sind diejenigen Merkmale im Sprachsignal, die primär etwas aussagen über die Beziehung zwischen Sprecher und Hörer oder aber über den Sprecher selbst, wie etwa seine Einstellung zum aktuell Gesagten. In diesem Fall haben die Intonationsmerkmale metakommunikative Funktion: Sie geben Auskunft, ob der Sprecher das aktuell Gesagte als ernst, lustig, traurig oder vielleicht auch als angsterregend einstuft. Neben den paralinguistischen Merkmalen mit metakommunikativer Funktion gibt es noch eine weitere Subklasse, nämlich diejenigen Merkmale, die auf den Charakter oder die Gemütsverfassung des Sprechers schließen lassen, also ob eine Stimme kindisch oder erfahren, freundlich oder schlecht gelaunt, fröhlich oder melancholisch klingt. Diese Merkmale sind es, die in weiten Teilen dasjenige ausmachen, was allgemein als „Persona“ bezeichnet wird, und auch sie müssen natürlich über den gesamten Dialog-Verlauf hinweg konsistent sein. Allerdings muss der Sprecher hier nicht, anders als bei den metakommunikativen Merkmalen, den bisherigen Gesprächsverlauf kennen.

In manchen Fällen liefert auch der Text eines einzelnen Prompts klare Indikatoren, welche metakommunikative Intonationsstruktur zu wählen ist, beispielsweise durch das Vorkommen von Wörtern wie „leider“. Der Sprecher erkennt dadurch sofort, dass eine ungünstige Situation dargelegt werden muss und dass sich dafür eine ernsthafte oder zumindest sachliche Intonation empfiehlt. In vielen Fällen jedoch stellt der Text keine klaren Indikatoren zu Verfügung:

- (10) System Als nächstes brauche ich den Vornamen der Person. <2 Sekunden Pause> Kennen Sie den Vornamen nicht, sagen Sie einfach ‚weiter‘.
- User Horst
- System Gut. Und in welcher Straße wohnt die Person?

Vor allem die Intonation des Diskursmarkers „gut“ ist hier wichtig. Solche Textelemente variieren mit der Intonation stark in ihrer kommunikativen Funktion. Gleiches gilt auch für viele der sogenannten Interjektionen wie „oh“ oder „hmm“: je nach Betonung kann zum Beispiel „hmm“ Zustimmung, Ablehnung oder Nachdenklichkeit signalisieren. Wie eingangs beschrieben, können Diskursmarker wie „gut“ Rückmeldung über die vom User getätigte Eingabe geben. In einem Zusammenhang wie in Beispiel 10 bedeutet „gut“ etwa soviel wie: „Ihre letzte Angabe wurde von mir registriert. Sie hat einen Teil zum Gelingen des gesamten Dialogs beigetragen.“. Die Intonation von „gut“ sollte daher Optimismus signalisieren.

- (11) System Als nächstes brauche ich den Vornamen der Person. <2 Sekunden Pause> Kennen Sie den Vornamen nicht, sagen Sie einfach ‚weiter‘.
- User ‚weiter‘.
- System Gut. Und in welcher Straße wohnt die Person?

In diesem Beispiel trägt der User mit seiner Angabe nicht dazu bei, dass dem erfolgreichen Ende des Dialogs ein Schritt näher gekommen wird. Das „gut“ bedeutet daher in diesem Zusammenhang so viel wie: „Ihre letzte Angabe wurde von mir registriert. Ich erhöhe meine Konzentration, um die jetzt noch verbleibenden Möglichkeiten auszuschöpfen.“. Eine Intonation wie unter 10 würde daher nicht zum bisherigen Gesprächsverlauf passen. In Beispiel 11 sollte „gut“ vielmehr mit nachdenklich-konzentrierter Stimme gesprochen werden.

In der Praxis werden in Fällen wie diesem häufig keine Unterschiede gemacht. Entweder das „gut“ wird nur in einer Intonationsvariante abgespielt, unabhängig vom Kontext, oder aber es wird komplett weggelassen. Das Fehlen von Diskursmarkern hat jedoch, wie eingangs gezeigt, meistens den Effekt, dass das System so wirkt, als wäre es nicht wirklich am Gespräch beteiligt, was eine geringere Nähe zwischen User und System zur Folge hat. Es sollten also mehrere Versionen des Prompts aufgenommen werden, eine optimistische und eine nachdenklich-konzentrierte, die dann in Abhängigkeit vom User-Input ausgegeben werden. Die jeweilige Art der Intonation zieht sich zwar durch den gesamten Prompt, der Schwerpunkt liegt allerdings auf dem jeweiligen Diskursmarker. Es empfiehlt sich daher, den Diskursmarker vom Rest des Textes abzuspalten, also diesen Text auf zwei Prompts aufzuteilen. Eine Audiofile-Konkatenation ist dabei unproblematisch, da nach „gut“ eine kurze Pause erfolgt. Auf diese Weise ergeben sich verschiedene Vorteile: Der Sprecher muss nicht über die verschiedenen Kontexte der Textpassage aufgeklärt werden, sondern spricht den neutralen Text (ohne Diskursmarker) und wird anschließend lediglich angehalten, das Wort „gut“ in verschiedenen Intonationsvarianten zu sprechen. Zusätzlich können noch weitere Diskursmarker vorgesehen werden, wie zum Beispiel „mmh“ oder „okay“, einmal in optimistischer und einmal in nachdenklich-konzentrierter Variante, verbunden mit einem

Randomized Prompting. Dadurch erreicht man zusätzlich eine sprachliche Abwechslung, die das System lebendiger und natürlicher wirken lässt und insbesondere bei High-Touch-Anwendungen von den Usern erfahrungsgemäß sehr geschätzt wird.

Fazit

Adäquate Intonierungen sind für eine gelungene und flüssige Kommunikation essentiell. Sie geben Aufschluss darüber, wie das Gesagte einzuordnen ist, und zwar wie es im bisherigen Gesprächsverlauf einzuordnen ist und wie der Sprecher zu dem Gesagten steht. Anders formuliert: Die Intonation gibt Aufschluss darüber, wie das Gesagte gemeint ist. Zudem helfen Intonationsstrukturen dabei, die Struktur komplexer Texte herauszuarbeiten und damit leichter verständlich zu machen.

Um die adäquate Intonation zu treffen, muss der Sprecher in den meisten Fällen den Text zumindest ein Stückweit in seinem Zusammenhang verstehen. Ist der Designer während des Recordings anwesend, kann er den Sprecher jeweils auf die gewünschte Intonationsstruktur hinweisen. Ist der Designer nicht anwesend, kann durch beschriebene Notations- und Design-Methoden nachgeholfen werden.

Inadäquate Intonierungen haben den Effekt, dass die Persona einer Anwendung „enttarnt“ wird. Dieses Phänomen ist vor allem vom Einsatz ungetunter TTSen bei längeren Textpassagen bekannt: Im Moment inadäquater Intonation wird indirekt deutlich, dass das System nicht wirklich versteht, und das kann aus der intuitiven Sicht des Users nur bedeuten, dass das System begrenzte Intelligenz und damit eine begrenzte Gesamtleistung besitzt. Adäquate Intonationsstrukturen hingegen tragen dazu bei, dass eine Anwendung lebhafter und dem User zugewandter erscheint. Auf diese Weise kann eine Nähe zwischen User und System entstehen, was für die subjektiv empfundene Systemleistung bekanntermaßen maßgeblich ist.